

Bioinformática

El ADN a un solo clic



David Roldán Martínez



Ra-Ma®

Bioinformática

El ADN a un solo clic



Bioinformática

El ADN a un solo clic

David Roldán Martínez





Bioinformática El ADN a un solo clic
© David Roldán Martínez

© De la Edición Original en papel publicada por Editorial RA-MA
ISBN de Edición en Papel: 978-84-9964-528-5
Todos los derechos reservados © RA-MA, S.A. Editorial y Publicaciones, Madrid, España.

MARCAS COMERCIALES. Las designaciones utilizadas por las empresas para distinguir sus productos (hardware, software, sistemas operativos, etc.) suelen ser marcas registradas. RA-MA ha intentado a lo largo de este libro distinguir las marcas comerciales de los términos descriptivos, siguiendo el estilo que utiliza el fabricante, sin intención de infringir la marca y solo en beneficio del propietario de la misma. Los datos de los ejemplos y pantallas son ficticios a no ser que se especifique lo contrario.

RA-MA es una marca comercial registrada.

Se ha puesto el máximo empeño en ofrecer al lector una información completa y precisa. Sin embargo, RA-MA Editorial no asume ninguna responsabilidad derivada de su uso ni tampoco de cualquier violación de patentes ni otros derechos de terceras partes que pudieran ocurrir. Esta publicación tiene por objeto proporcionar unos conocimientos precisos y acreditados sobre el tema tratado. Su venta no supone para el editor ninguna forma de asistencia legal, administrativa o de ningún otro tipo. En caso de precisarse asesoría legal u otra forma de ayuda experta, deben buscarse los servicios de un profesional competente.

Reservados todos los derechos de publicación en cualquier idioma.

Según lo dispuesto en el Código Penal vigente ninguna parte de este libro puede ser reproducida, grabada en sistema de almacenamiento o transmitida en forma alguna ni por cualquier procedimiento, ya sea electrónico, mecánico, reprográfico, magnético o cualquier otro sin autorización previa y por escrito de RA-MA; su contenido está protegido por la Ley vigente que establece penas de prisión y/o multas a quienes, intencionadamente, reprodujeran o plagiaran, en todo o en parte, una obra literaria, artística o científica.

Editado por:
RA-MA, S.A. Editorial y Publicaciones
Calle Jarama, 33, Polígono Industrial IGARSA
28860 PARACUELLOS DE JARAMA, Madrid
Teléfono: 91 658 42 80
Fax: 91 662 81 39
Correo electrónico: editorial@ra-ma.com
Internet: www.ra-ma.es y www.ra-ma.com

Maquetación y diseño portada: Antonio García Tomé

ISBN: 978-84-9964-575-9

E-Book desarrollado en España en Octubre de 2015

*A todos aquellos que se empeñaron en hacerme
caer una y otra vez, porque el levantarme
me ha hecho ser más fuerte.
A todos aquellos que intentaron acobardarme, porque
el temple que adquirí me ha hecho perder el miedo.
A todos aquellos que me infundieron desánimo,
porque superarlo ha hecho crecer mi autoestima.
A todos aquellos que me crearon problemas, porque
solucionarlos me ha hecho más inteligente y a todos
aquellos que amaron y me aman, que me ayudaron y
me ayudan porque sin ellos no sería nada.*

ÍNDICE

AUTOR	11
PRÓLOGO	13
CAPÍTULO 1. INTRODUCCIÓN	15
1.1 A QUIÉN VA DESTINADO ESTE LIBRO	16
1.2 ESTRUCTURA DE ESTE LIBRO	16
1.3 LEYENDAS	17
CAPÍTULO 2. FUNDAMENTOS BIOLÓGICOS	19
2.1 FISIOLOGÍA CELULAR	20
2.2 MORFOLOGÍA DEL CROMOSOMA	22
2.3 ÁCIDOS NUCLEICOS	24
2.3.1 ADN	24
2.3.2 ARN	25
2.3.3 Código genético	26
2.4 DOGMA CENTRAL DE LA BIOLOGÍA MOLECULAR	27
2.5 REGULACIÓN GÉNICA	29
CAPÍTULO 3. FORMATOS DE FICHEROS	31
3.1 DATOS EN BRUTO	32
3.2 FASTA	34
3.3 FASTAQ	35
3.4 SAM/BAM	36
3.5 GFF/GFF3	41
3.6 GVF	43

3.7	VCF.....	45
3.8	BED.....	48
CAPÍTULO 4. BASES DE DATOS GENÓMICAS.....		51
4.1	¿QUÉ ES UNA BASE DE DATOS GENÓMICA?.....	54
4.2	CLASIFICACIÓN DE LAS BASES DE DATOS GENÓMICAS.....	54
4.3	CARACTERÍSTICAS DE LA INFORMACIÓN GENÓMICA.....	56
4.4	CONSTRUCCIÓN DE UNA BASE DE DATOS GENÓMICA.....	58
4.5	MODELADO DE INFORMACIÓN GENÓMICA.....	59
4.6	INTEGRACIÓN DE BASES DE DATOS BIOLÓGICAS	60
CAPÍTULO 5. PRÁCTICA 1: DISEÑO DE BASES DE DATOS BIOLÓGICAS.....		63
5.1	DISEÑO RELACIONAL	63
5.2	DISEÑO XML	67
CAPÍTULO 6. PRINCIPALES BASES DE DATOS GENÓMICAS.....		73
6.1	GENBANK.....	73
6.1.1	Formato del registro	75
6.1.2	Cabecera	78
6.1.3	Sección de características.....	81
6.1.4	Sección ORIGIN	82
6.2	REFSEQ.....	82
6.3	UNIPROT	84
6.4	PDB.....	84
6.4.1	Formato del registro	85
6.4.2	Tipos de registros	85
6.4.3	Estructura del fichero	88
6.5	OTRAS BASES DE DATOS GENÓMICAS	90
6.5.1	Bases de datos de secuencias de ADN	90
6.5.2	Bases de datos de secuencias de ARN	92
6.5.3	Bases de datos de secuencias de proteínas	93
6.5.4	Bases de datos de patrones y perfiles	94
6.5.5	Bases de datos clínico-genéticas	95
6.5.6	Bases de datos de mutaciones y SNP	96
6.5.7	Bases de datos de genómica funcional.....	96
CAPÍTULO 7. PRÁCTICA 2: BÚSQUEDA DE SECUENCIAS.....		97
7.1	SECUENCIAS DE ORGANISMOS PROCARIOTAS.....	97
7.2	SECUENCIAS DE ORGANISMOS EUCARIOTAS	103
7.3	BÚSQUEDA DE VARIACIONES	106
7.4	EJEMPLO DE ESTUDIO DE UNA PROTEÍNA	108

CAPÍTULO 8. ANÁLISIS DE SECUENCIAS	113
8.1 DETECCIÓN DE ORF	114
8.2 ANÁLISIS DE CALIDAD	115
8.3 ALINEAMIENTO	115
8.3.1 Gráficos de puntos	117
8.3.2 Alineamiento de pares	118
8.3.3 Alineamiento múltiple	119
8.3.4 Puntuación del alineamiento	120
8.4 IDENTIFICACIÓN DE VARIACIONES	122
8.5 ANOTACIÓN	126
8.6 VISUALIZACIÓN	126
8.7 PIPELINES ANALÍTICOS Y SISTEMAS DE FLUJO DE TRABAJO	127
CAPÍTULO 9. PRÁCTICA 3: ANÁLISIS DE SECUENCIAS	129
9.1 ANÁLISIS DE LA CALIDAD CON VECSCREEN	129
9.2 ANÁLISIS DE LA COMPOSICIÓN DEL ADN	135
9.2.1 Búsqueda de palabras	135
9.2.2 Estadísticas de la secuencia con Genomatix	137
9.2.3 Búsqueda de repeticiones	139
9.2.4 Búsqueda de ORF	150
9.3 ALINEAMIENTO DE SECUENCIAS CON BLASTN	160
9.4 EDICIÓN DE ALINEAMIENTOS	166
9.4.1 Creación de grupos	175
9.4.2 Reordenación del alineamiento	177
9.4.3 Adición y borrado de huecos	178
9.5 BÚSQUEDA DE SECUENCIAS HOMÓLOGAS CON SIB-BLAST	179
9.6 ALINEAMIENTO MÚLTIPLE	183
9.6.1 Alineamiento múltiple con Clustal Omega	185
9.6.2 Alineamiento múltiple con MUSCLE	192
9.6.3 Alineamiento múltiple con T-Coffee	193
CAPÍTULO 10. PROTEÓMICA	195
10.1 GENERALIDADES	195
10.2 ESTRUCTURA DE LAS PROTEÍNAS	196
10.3 MÉTODOS DE PREDICCIÓN	197
10.4 MODELADO POR HOMOLOGÍA	198
10.5 RECONOCIMIENTO DE PLIEGUES	199
CAPÍTULO 11. PRÁCTICA 4: ANÁLISIS DE PROTEÍNAS	201
11.1 ANÁLISIS BLAST	201
11.2 BÚSQUEDA DE DOMINIOS FUNCIONALES	207

11.2.1	Búsqueda de dominios con EBI-Interpro	208
11.2.2	Búsqueda de dominios con PFAM	211
11.3	PREDICCIÓN DE LA UBICACIÓN SUBCELULAR.....	214
11.4	BÚSQUEDA DE ESTRUCTURAS DE REFERENCIA	216
11.5	BÚSQUEDA DE MOTIVOS	221
11.6	ANÁLISIS DE LA ESTRUCTURA PRIMARIA DE UNA PROTEÍNA	237
11.6.1	Traducción del ADN en secuencia proteica	237
11.6.2	Predicción de las propiedades físico-químicas.....	239
11.7	PREDICCIÓN DE LA ESTRUCTURA SECUNDARIA.....	242
11.8	PREDICCIÓN DE LA ESTRUCTURA TERCIARIA.....	247
11.9	PREDICCIÓN DE GENES CON GENSCAN	249
BIBLIOGRAFÍA.....		255
ÍNDICE ALFABÉTICO		259



AUTOR

David Roldán Martínez es doctor ingeniero de Telecomunicación y máster en Redes Corporativas e Integración de Sistemas por la Universidad Politécnica de Valencia (UPV). Ha trabajado en empresas de consultoría y desarrollo de proyectos y productos relacionados con las tecnologías de la información y hoy en día es analista de aplicaciones del ASIC de la UPV.

Ha contribuido activamente en comunidades *open source* como Sakai (software *e-learning* en donde desempeñó diversos cargos de responsabilidad mundial y que le otorgó en 2011 el *Sakai Fellow Award* por su participación activa en el desarrollo y puesta en marcha de Sakai alrededor del mundo). Actualmente, es desarrollador y responsable de i18n de Jalview, una de las herramientas para la gestión y anotación de secuencias genómicas más utilizadas a nivel mundial.

Además, ha sido profesor del Departamento de Comunicaciones de la Universidad Politécnica de Valencia y ha impartido formación de posgrado en distintas universidades e instituciones. Es miembro de ACTA (Asociación de Autores Científicos y Técnicos) y, siempre preocupado por la divulgación científico-tecnológica, dispone en su haber de numerosos libros y artículos relacionados con diversos aspectos de las TIC.

AGRADECIMIENTOS

Este libro es especial para mí por muchas razones. Supone haber cumplido una ilusión que tenía desde hace mucho tiempo: transmitir mi pasión por la Biología. Por este motivo quisiera agradecer, en primer lugar, a RA-MA el haber confiado en mí y aceptado mi propuesta de proyecto.

Y, cómo no, muchas gracias a mi mujer y a mis hijos por el tiempo que les he robado, que no tiene precio. Espero que el sacrificio haya merecido la pena.

No me olvidaré tampoco de Óscar Pastor, Ana Ciudad y Mercedes Fernández que, conscientemente a veces e inconscientemente otras, me han animado a alcanzar esta meta.

También me gustaría una mención especial a Luis Ruano, mi profesor de Biología en 3.º BUP en el Instituto José Conde García de Almansa (Albacete) y que, en su día, supo sembrar en mí la pasión por este mundo que, aunque ha tardado en germinar, finalmente lo ha hecho.

No me dejaré a Diana, Mar, Richard, María José, Jacobo, Juan Vicente, Pepe, Manolo, Javier, Esteban, Sebastián, Aurora, David, Sergio, Tanja y Benjamín por su ayuda y sus opiniones.

PRÓLOGO

Durante el siglo pasado y a principios de este, asistimos a la revolución de las Tecnologías de la Información y las Telecomunicaciones. En apenas una década pasamos de buscar información en enormes tomos de papel que debían consultarse por turnos con otros usuarios, en bibliotecas en donde, con suerte tras varias horas la búsqueda tenía éxito, a disponer de amplios catálogos bibliográficos en línea, ubicados en cualquier parte del planeta y cuyas consultas pueden efectuarse sentado cómodamente en una hamaca de la playa desde el teléfono móvil o la tableta.

El gran reto de estas dos próximas décadas será, sin duda alguna, la revolución biológica, en cuanto prácticamente cualquier tipo de experimento se plantea en un contexto genómico. En efecto, el hecho de que los avances en las tecnologías de secuenciación permitan el análisis no solo de la secuencia de ADN de un individuo sino también de su metaboloma con un nivel de detalle suficiente como para predecir la evolución de una enfermedad o una terapia, va a suponer una profunda transformación conceptual, técnica y tecnológica de la Biología en general y de la Medicina en particular, que sentará las bases de la Medicina Predictiva, Personalizada, Preventiva y Participativa.

Por otra parte, el creciente volumen de información heterogénea en origen y significado, a lo que se suma su complejidad y la necesidad de analizarla en distintos contextos ha convertido a la Bioinformática en una pieza clave en el futuro cercano de la humanidad.

A través de esta obra el autor, David Roldán, intenta darnos una visión eminentemente práctica de los principios básicos de la Bioinformática, explicando

con detalle los aspectos fundamentales de esta rama híbrida entre Computación y Biología.

Desde estas líneas aprovecho la oportunidad para darle mi más sincera enhorabuena por un trabajo tan conseguido, así como para felicitarle porque, sinceramente, pienso que el esfuerzo y la ilusión que ha invertido han merecido la pena.

Óscar Pastor López

Catedrático de universidad

Director del grupo de investigación PROS

1

INTRODUCCIÓN

La información biológica se produce a un ritmo fenomenal, hasta tal punto que se estima que cada 15 meses se duplica su tamaño. El resultado, además de una enorme montaña de datos biológicos, es que se hace imprescindible la utilización de ordenadores que asistan a los científicos en la gestión de dicha información. Y es en este contexto en donde se enmarca la Bioinformática. Aunque son muchas las definiciones que se han dado de la Bioinformática a lo largo de los últimos años, todas coinciden en señalar dos aspectos básicos.

En primer lugar, la Bioinformática se encarga de la organización de la información biológica de manera que se simplifique y optimice el acceso a la misma por parte de los usuarios (investigadores, científicos y otras aplicaciones) así como de la actualización constante de dicha información a medida que se vaya produciendo.

El segundo objetivo es proporcionar a los usuarios las herramientas y recursos necesarios para analizar los datos biológicos. Efectivamente, no solamente se requiere poder acceder a la información sino también herramientas que permitan realizar búsquedas en grandes volúmenes de datos de manera sencilla, o cruzar información de distintas fuentes bien para contrastarla o bien incluso para generar información derivada. Este campo de la minería de datos ofrece una proyección de futuro realmente espectacular.

Este libro aborda el estudio de la Bioinformática centrándose, precisamente, en estos dos enfoques. En los primeros capítulos se estudian las características de la información biológica y qué principios es necesario tener en cuenta a la hora de diseñar un sistema de información biológico. Por otro lado, en su segunda parte, la obra se centra en presentar herramientas y métodos de análisis de dicha información biológica. Es importante subrayar que se ha hecho especial hincapié en seleccionar

herramientas de fuentes abiertas (*open source*), ya que de esta manera estarán al alcance de cualquier lector, sin tener que depender de onerosos costes de licencias.

Además, para reforzar esta visión, se presenta un conjunto de ejercicios y cuestiones prácticas, donde se proponen y resuelven diferentes ejercicios que permiten profundizar en los conceptos teóricos introducidos.

1.1 A QUIÉN VA DESTINADO ESTE LIBRO

El libro está orientado tanto a informáticos, como a biólogos y estudiantes de los primeros cursos de Biotecnología, Ingeniería Biomédica y carreras afines pues aborda la temática desde un punto de vista práctico planteando, cuando la temática lo permite, ejercicios y cuestiones resueltas, con una extensión amplia en los comentarios de las soluciones.

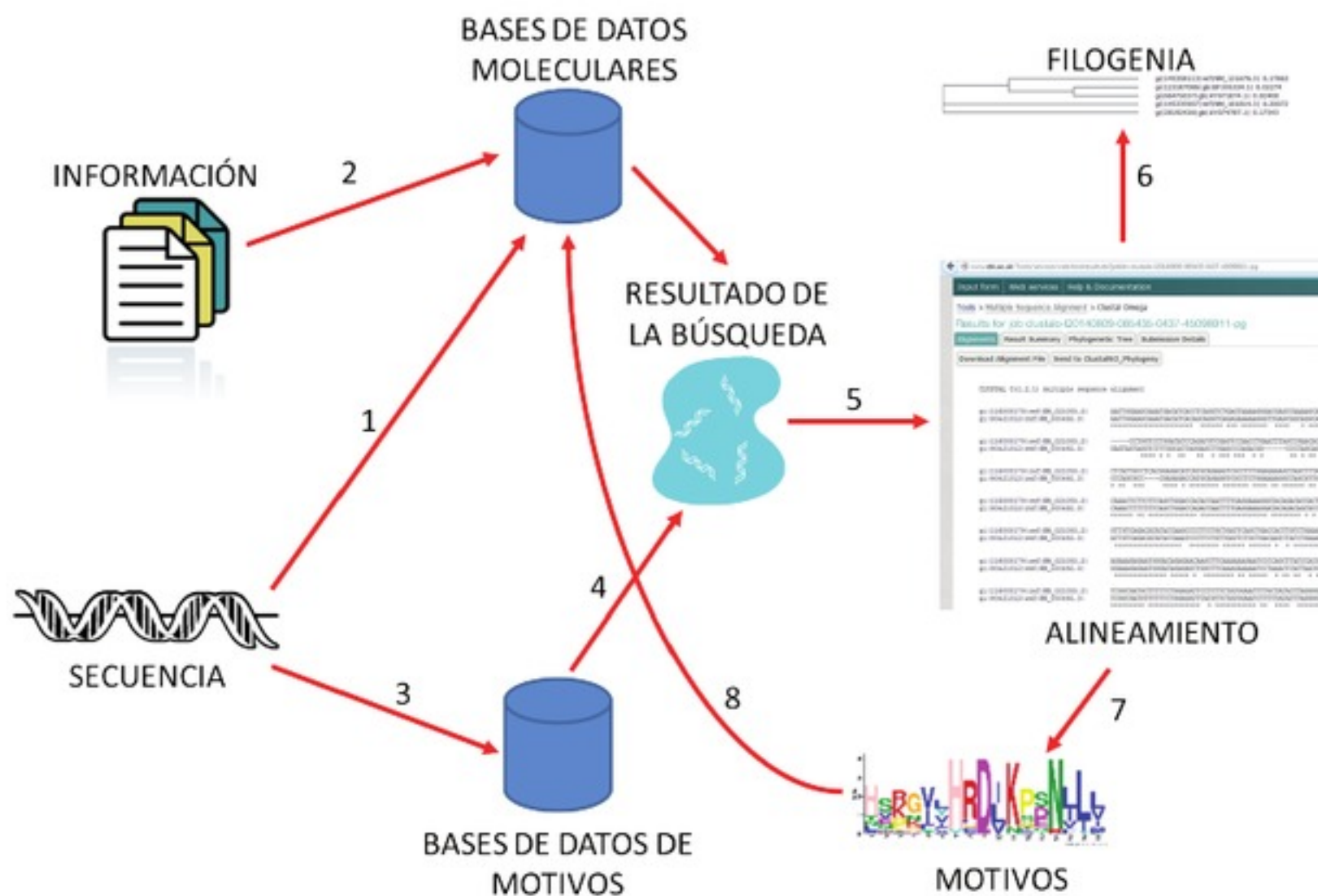
Si el lector es un usuario con amplia experiencia en Bioinformática, se verá tentado a saltarse parte del contenido de este libro. La experiencia demuestra que muchos de los conceptos básicos han sido pasados por alto incluso por usuarios experimentados. Para evitar esto, en cada capítulo del libro se resaltan con leyendas gráficas (presentadas en la sección 1.3) aspectos importantes que hay que tener en cuenta, así como consejos prácticos.

1.2 ESTRUCTURA DE ESTE LIBRO

A lo largo del libro, y sobre todo en la segunda parte del mismo, se hará uso del siguiente esquema para las prácticas y para los ejemplos presentados.

El uso de la Bioinformática en la Biología Molecular se ha visto acelerado por la capacidad del software para analizar datos de secuencias así como por la posibilidad de almacenar conocimiento sobre las mismas en bases de datos de tal manera que, cuando un biólogo se enfrenta al problema de identificar una determinada secuencia con el fin de averiguar, por ejemplo, qué función o qué estructura tendrá, se realiza una comparación de dicha secuencia con las guardadas en distintas bases de datos. En esta comparación y posterior procesado se identifica una serie de pasos bien diferenciados (ver Figura 1.1). En primer lugar, se realiza una búsqueda en bases de datos genómicas para determinar si existe algún conocimiento anterior de dicha secuencia (paso 1), de secuencias similares (paso 2) o ciertas regiones de la secuencia (motivo) con significado funcional o estructural (paso 3). Los resultados de

la búsqueda (paso 4) se alinean con la secuencia bajo estudio (paso 5) para obtener, por una parte, el árbol filogenético (paso 6) y, por otra, zonas conservadas (paso 7). A partir de las zonas conservadas es posible definir motivos (paso 8) e iniciar el proceso de nuevo. Sobre este proceso básico, es posible construir otros más complejos como la predicción de genes o la predicción de la estructura de secuencias de proteínas.



Los 11 capítulos de este libro pretenden que el lector adquiera los conocimientos básicos de Bioinformática que le permitan desenvolverse con cierta soltura cuando se le plantee resolver el problema de averiguar la mayor cantidad de información posible sobre una determinada secuencia.

1.3 LEYENDAS

Como ya se ha comentado, en cada capítulo de este libro se destacan algunos aspectos importantes que hay que tener en cuenta y se dan también determinadas recomendaciones prácticas. Para destacar estos puntos se utilizan las leyendas gráficas que se presentan a continuación.



Aspecto que hay que tener en cuenta



Recomendación práctica



Prueba tú mismo



Ejemplo o punto teórico desarrollado en prácticas



Recuerda

FUNDAMENTOS BIOLÓGICOS

La Genética, en su más amplio sentido, es la rama de la ciencia encargada de estudiar el funcionamiento y la transmisión de los genes, responsables últimos de la herencia.

Sus bases fueron establecidas por el monje austriaco Gregor Mendel en 1866 con su trabajo sobre la hibridación de los guisantes. Tradicionalmente, se ha basado en los procesos de mutación y selección. Los experimentos de Mendel concluyeron que la herencia reside en unas unidades discretas que pasan de generación en generación de manera independiente. Estas unidades, a las que se acuñó con el nombre de *elemente*, en 1909 tomarían el nombre de genes. La rama de la Genética encargada de la transmisión de los caracteres hereditarios de una generación a otra es la Genética Clásica, y está relacionada con la Genética Poblacional que, basándose en la genética de familias individuales, trata de extrapolar las conclusiones a grupos de individuos más grandes.

Ya en la época moderna se estableció la Genética Molecular, encargada del estudio físico-químico del ADN, el ARN y las proteínas. También se conoce con el nombre de Genómica y es en la que nos centraremos en este libro.

Finalmente, encontramos la Genética Cuantitativa. Se trata de un campo extremadamente matemático cuyo objetivo es el análisis de relaciones estadísticas entre los genes y los rasgos que imprimen.



Este capítulo introduce, muy brevemente, los fundamentos elementales de Biología para comprender el contexto en que se enmarca todo el desarrollo posterior.

Algunos conceptos se han simplificado con el fin de facilitar su comprensión por parte del lector no iniciado, aun a riesgo de ser poco estrictos. Rogamos a los expertos que sepan disculpar esta aparente falta de rigurosidad.

2.1 FISIOLÓGÍA CELULAR

Una célula puede definirse como la unidad mínima capaz de realizar todas las funciones asociadas a un organismo vivo (nutrición, relación y reproducción).

A la hora de clasificar las células, existen organismos unicelulares y organismos pluricelulares.

Desde el punto de vista unicelular, podemos distinguir entre:

- Células procariotas: carecen de núcleo y otros orgánulos (vacuolas y mitocondrias). En este caso, el material genético está disperso en una o más regiones nucleoides sin la protección de una membrana nuclear. A este grupo pertenecen las bacterias y algunas algas.
 - Células eucariotas: se caracterizan porque tienen un núcleo en el que se localiza el ADN y que almacena el material genético de la célula. Además, en el citoplasma existente entre el núcleo y la membrana celular que marca los límites de la célula, encontramos numerosos orgánulos con funciones muy definidas.
- La mayoría de los eucariotas son organismos pluricelulares, aunque también los hay unicelulares como las levaduras.
- Virus y organismos subvirásicos: estrictamente hablando, no son seres celulares, ya que no son capaces de sobrevivir de manera independiente y necesitan de otra célula a la que infectan. Tienen una estructura muy simple formada por un filamento de ADN o ARN (nunca los dos en el mismo virus), envuelto por una cápsula proteica especial.

No obstante, en la naturaleza resulta muy común encontrar agrupaciones de células especializadas denominadas tejidos y que son distintos en los animales y en las plantas.

La célula tiene un alto nivel de organización sorprendentemente complejo. En la Tabla 2.1 se recogen los componentes que forman parte de la estructura celular de organismos eucariotas y procariotas.

Estructura		Descripción	Función
Núcleo celular	Nucleoplasma	Rodeado por una doble membrana, contiene el nucléolo y los cromosomas	Control de la célula
	Nuécleo	Contiene ARN y proteínas	Biosíntesis de ribosomas
	Cromatina	ADN nuclear	Contiene los genes
Sistema de membranas	Membrana plasmática	Contiene el citoplasma	Establece los límites de la célula viva con el entorno que la rodea, regula el intercambio de sustancias con dicho entorno, mantiene la estructura celular y comunica la célula con otras
	Reticulo endoplasmático (RE) liso	Red de membranas internas	Biosíntesis de lípidos
	Reticulo endoplasmático (RE) rugoso		Biosíntesis de proteínas
Citoplasma	Ribosomas	Gránulos unidos al RE o libres por el citoplasma	Biosíntesis de proteínas
	Aparato de Golgi	Conjunto de sáculos membranosos	Modifica, empaqueta y distribuye proteínas a los orgánulos de la célula
	Lisosomas	Sáculos membranosos (en animales)	Contienen enzimas que participan en los procesos de nutrición celular
	Vacuolas	Sáculos membranosos (plantas, hongos y algas)	Participan en la nutrición celular
	Microcuerpos	Sáculos membranosos	Contiene enzimas para el metabolismo celular
	Mitocondrias	Sáculos membranosos	Se encarga de la respiración celular y de la biosíntesis de energía
	Plastos	Sistemas membranosos (plantas y hongos)	Respiración celular
Citoesqueleto	Microtúbulos	Túbulos huecos	Funciones estructurales
	Microfilamentos	Estructuras sólidas y cilíndricas	Funciones estructurales y movimiento celular
	Centriolos	Cilindros huecos	Participan en la división celular
	Cilios	Túbulos	Movimiento celular
	Flagelos	Microtúbulos	Movimiento celular

Tabla 2.1. Estructura de la célula eucariota

2.2 MORFOLOGÍA DEL CROMOSOMA

Los cromosomas son pequeños bastones en que organiza la cromatina del núcleo durante la división celular y que están compuestos de ADN. La transmisión del fenotipo exige que se copie el material genético de una generación a otra. Este proceso puede ser relativamente simple, como en algunas células procariotas que tienen un único cromosoma circular que se transmite de padres a hijos durante la mitosis, o mucho más complejo, como ocurre en las células eucariotas en la mitosis y la meiosis.



El ADN que contiene los genes de la célula se encuentra en los cromosomas.

El número de cromosomas es característico de cada especie (ver Tabla 2.2). Por ejemplo, en el ser humano (*Homo sapiens*) el cariotipo está formado por 46 cromosomas. Los cromosomas se dividen en sexuales (determinan el sexo del sujeto) y autosómicos (no sexuales). En el caso del ser humano, existen 2 cromosomas sexuales y 44 autosómicos. Se denomina genoma al conjunto de genes de un organismo.

Especie	Número cromosómico
Ser humano	46 (23 pares)
Bovinos	60 (30 pares)
Oveja	54 (27 pares)
Maíz	20 (10 pares)
Tabaco	48 (24 pares)
Tomate	24 (12 pares)

Tabla 2.2. Número de cromosomas de algunas especies

Los cromosomas suelen representarse como una X (ver Figura 2.1), aunque únicamente adoptan esta forma durante la división celular. La parte central del cromosoma, donde se cruzan sus brazos, recibe el nombre de centrómero, mientras que los extremos son los telómeros. Es, precisamente, en los telómeros en donde se encuentra una mayor densidad de ADN.

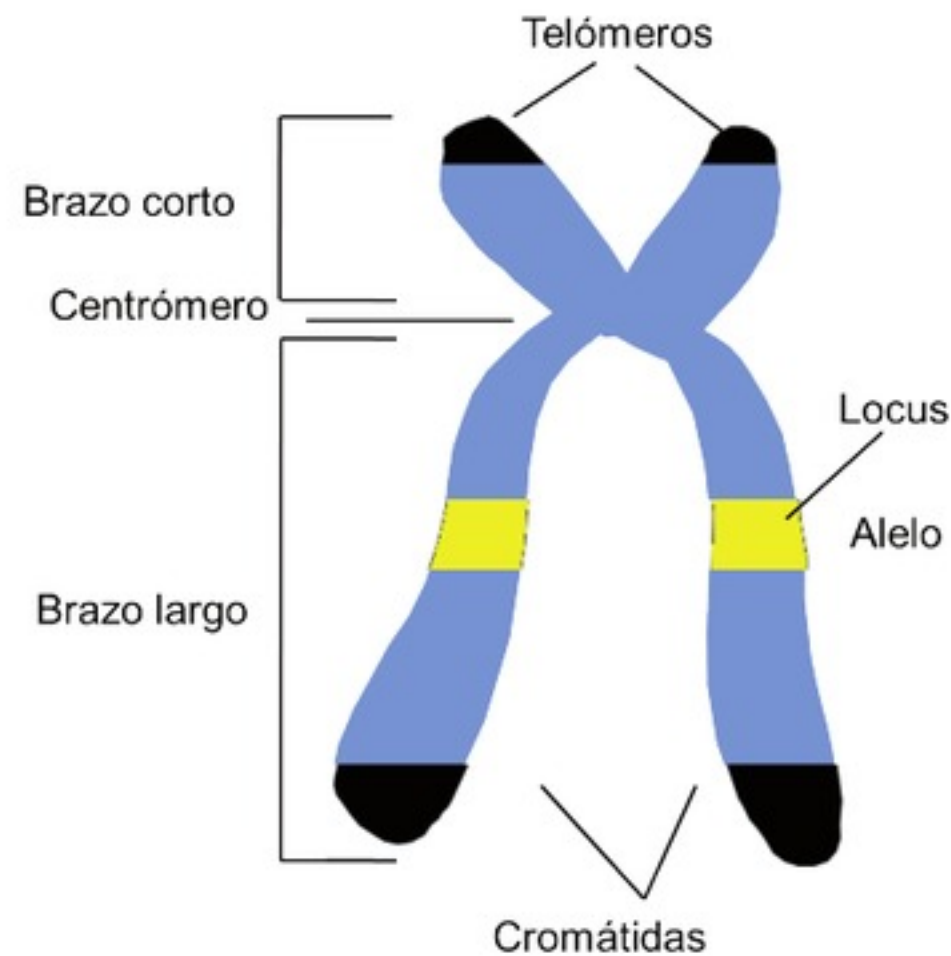


Figura 2.1. Morfología básica de un cromosoma eucariota

Los cromosomas son los portadores de los genes y cada cromosoma puede albergar más de un gen. Cada par de genes homólogos transporta la misma información genética, es decir, los mismos genes, aunque sus copias (también llamadas alelos) no necesariamente tienen que ser idénticas. La posición que ocupa un gen en un cromosoma dado recibe el nombre de locus.

A más bajo nivel, un gen es una secuencia de ADN que dicta las instrucciones para la síntesis de proteínas. Sin embargo, no todos los genes codifican proteínas. Algunos controlan el proceso de transcripción, tal y como veremos más adelante.

Por otra parte, en el cromosoma, además de ADN existen elementos reguladores transcripcionales y regiones conservadas (como la caja TATA de las eucariotas).



En el ADN de un cromosoma podemos encontrar genes, reguladores y regiones conservadas.

2.3 ÁCIDOS NUCLEICOS

Los ácidos nucleicos deben su importancia a que intervienen en el manejo de la información genética. En los seres vivos es posible encontrar dos tipos de ácidos nucleicos: el desoxirribonucleico (ADN) y el ribonucleico (ARN). Se trata de polímeros lineales compuestos por sucesiones de nucleótidos o bases nitrogenadas.

Las cinco bases se agrupan en bases púricas (si tienen una estructura en doble anillo) o bases pirimidínicas (si el anillo es simple). Las primeras son la adenina (A) y la guanina (G), mientras que las segundas son la timina (T), la citosina (C) y el uracilo (U). No todas las bases forman parte de los dos ácidos nucleicos. En el ADN encontramos adenina, guanina, timina y citosina, mientras que en el ARN solamente existen adenina, guanina, citosina y uracilo. La secuencia de bases nitrogenadas y la longitud de la cadena de nucleótidos es característica para cada especie viva. Sin embargo, los trabajos de Chargaff (1951) concluyeron que en todo ser vivo existe siempre la misma cantidad de adenina que de timina así como de citosina y guanina.

2.3.1 ADN

El ácido desoxirribonucleico o ADN es el portador de la información genética que pasará de generación en generación entre las células del organismo y que, como hemos visto anteriormente, se encuentra concentrado en los cromosomas.

La molécula de ADN tiene una forma de doble hélice en la que dos cadenas de bases complementarias se enrollan a lo largo de su eje central. Las bases y la longitud de la cadena de nucleótidos son característicos del organismo y se copian con exactitud durante la reproducción celular. Este tema será tratado detalladamente más adelante.

Todos los tipos de ADN están compuestos por las mismas cuatro bases y tienen estructura de doble hélice, si bien llevan a cabo funciones diferentes, que son:

- ADN nuclear: es el que se encuentra en el núcleo de la célula y responsable de la mayoría de funciones celulares. El ADN nuclear es el portador de los genes.
- ADN mitocondrial: los animales, las plantas y los hongos incorporan ADN en las mitocondrias (ADNmt) cuya misión es regular el metabolismo celular. Una curiosidad acerca del ADNmt es que, a diferencia de lo que ocurre con el ADN nuclear en donde la mitad del material genético

procede del padre y la otra mitad de la madre, todo el ADNmt se hereda de la madre.

- ADN de los cloroplastos: además de los dos anteriores, las plantas tienen unos orgánulos encargados de la fotosíntesis llamados cloroplastos que también contienen moléculas de ADN (ADNcp).

2.3.2 ARN

El ácido ribonucleico o ARN está compuesto por una sola cadena lineal de nucleótidos en la que el glúcido es la ribosa. La mayor parte de las células contienen de 2 a 8 veces más ARN que ADN.

Existen tres tipos de ARN, cuya síntesis se cataliza por tres tipos distintos de enzimas (ARN polimerasas), que son:

- ARN mensajero (ARNm): es una copia complementaria del fragmento del código genético del ADN. Su síntesis está catalizada por la ARN polimerasa II y suele tratarse de largas cadenas de nucleótidos.
- ARN de transferencia (ARNt): su síntesis está catalizada por la ARN polimerasa III. Es el más pequeño de los tres y se halla disperso por todo el citoplasma. Se conocen unos 50 tipos de ARNt, todos ellos tienen una configuración similar, con un brazo aceptor, en el que aparece siempre la secuencia CCA; y un anticodón, que es un triplete de bases nitrogenadas que determina el aminoácido que se va a unir a esa molécula de ARNt.
- ARN ribosómico (ARNr): está formado por moléculas muy largas y plegadas y su síntesis está catalizada por la ARN polimerasa



Hay experimentos que parecen confirmar que el ARN es la molécula a partir de la cual surgió la vida, pese a que actualmente sea el otro ácido nucleico (ADN) el portador de la información genética. Una de las propiedades más características del ARN es la autocatálisis, es decir, la capacidad de acelerar su propia formación.

2.3.3 Código genético

Cada aminoácido se corresponde con un conjunto de tres bases denominado codón y las secuencias de ADN y ARN se constituyen como tiras de estos tripletes o codones. Puesto que las bases que se combinan en el ADN son solamente 4 (A, G, C y T) y el número de nucleótidos por codón es tres, el número de codones teórico es de $4^3 = 64$. Dado que únicamente hay 20 aminoácidos posibles, se trata de un código degenerado, es decir, tiene múltiples símbolos para todos los aminoácidos, con excepción del Trp y la Met. La degeneración reside, generalmente, en la tercera posición de su codón (extremo 3'): el nucleótido de esta posición es mucho menos específico que el primero y que el segundo. Tres de las 64 secuencias no codifican ningún aminoácido sino que determinan el final de la cadena polipeptídica, por lo que se les conoce con el nombre de codones de terminación.

El conjunto de códigos posibles recibe el nombre de código genético y es universal. En la Tabla 2.3 se indica la codificación de bases de cada uno de los aminoácidos.

Ala (A)	GCU, GCC, GCA, GCG	Lys (K)	AAA, AAG
Arg (R)	CGU, CGC, CGA, CGG, AGA, AGG	Met (M)	AUG
Asn (N)	AAU, AAC	Phe (F)	UUU, UUC
Asp (D)	GAU, GAC	Pro (P)	CCU, CCC, CCA, CCG
Cys (C)	UGU, UGC	Sec (U)	UGA
Gln (Q)	CAA, CAG	Ser (S)	UCU, UCC, UCA, UCG, AGU, AGC
Glu (E)	GAA, GAG	Thr (T)	ACU, ACC, ACA, ACG
Gly (G)	GGU, GGC, GGA, GGG	Trp (W)	UGG
His (H)	CAU, CAC	Tyr (Y)	UAU, UAC
Ile (I)	AUU, AUC, AUA	Val (V)	GUU, GUC, GUA, GUG
Leu (L)	UUA, UUG, CUU, CUC, CUA, CUG		
Comienzo	AUG	Parada	UAG, UGA, UAA

Tabla 2.3. Código genético

2.4 DOGMA CENTRAL DE LA BIOLOGÍA MOLECULAR

La base química de la herencia es la molécula de ADN. En 1953 James Watson y Francis Crick postularon que la molécula de ADN está formada por largas cadenas de nucleótidos enrollados en espiral formando una doble hélice. Trabajos posteriores de Crick, Brenner, Nirenberg y Ochoa concluyeron la existencia de un mecanismo de transmisión de la herencia: la transcripción del mensaje genético del ADN al ARN y la síntesis de proteínas (ver Figura 2.2). Esto es lo que se conoce como Dogma central de la Biología Molecular.



Figura 2.2. Dogma central de la Biología Molecular

Aunque el proceso es esencialmente el mismo en células eucariotas y en procariotas, existen algunas diferencias pero quedan fuera del ámbito de este libro.

La Figura 2.3 esquematiza cómo se lleva a cabo el proceso completo:

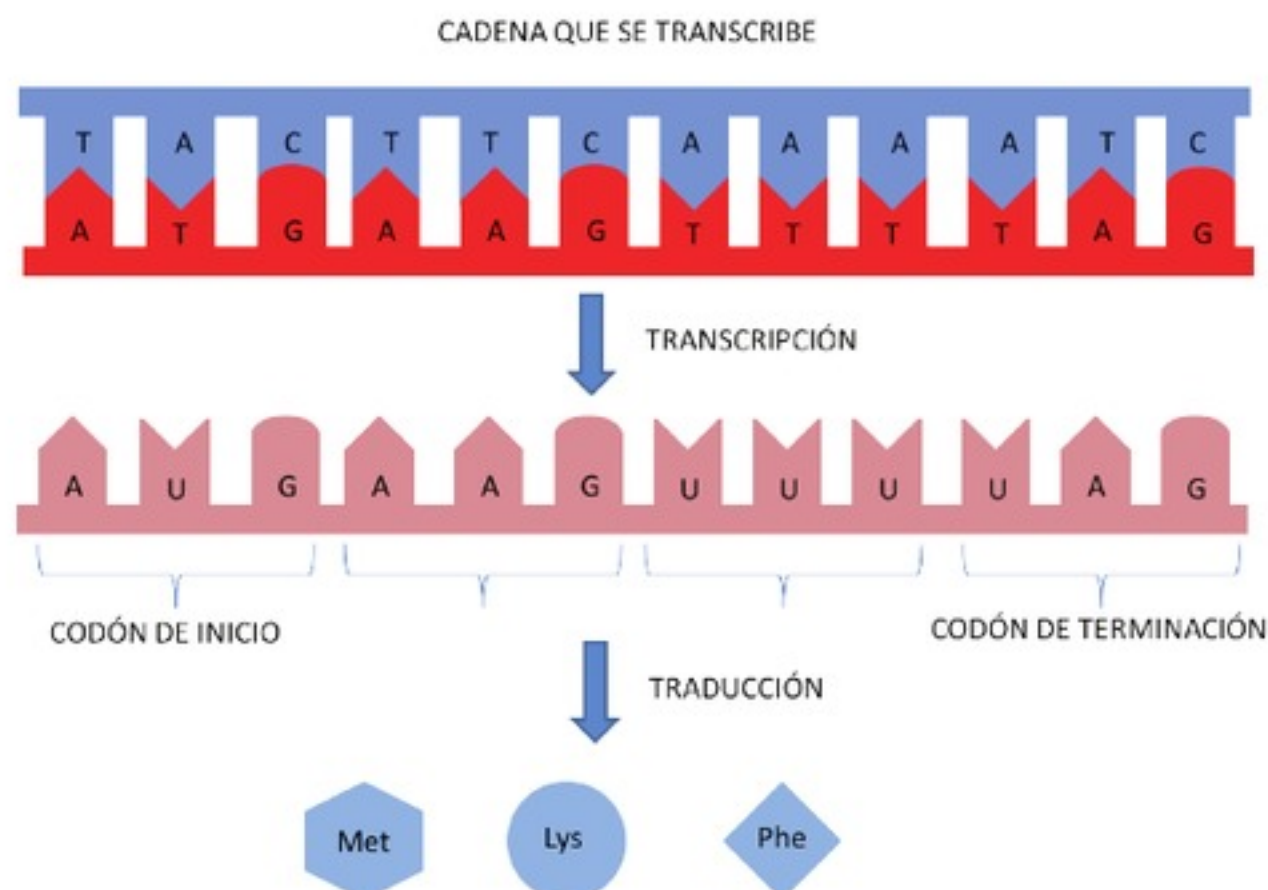


Figura 2.3. Procesos de transcripción y traducción

Durante la replicación de una cadena de ADN se obtendrán dos cadenas idénticas, una de las cuales irá a la célula madre y la otra a la hija. Comienza cuando se encuentra en la cadena original la secuencia de inicio, que marca el punto por el que debe abrirse la doble hélice y comenzar la replicación. Una vez abierta, se van ensamblando nucleótidos por complementariedad (transcripción), de manera que la hebra original se lee en sentido $3' \rightarrow 5'$ y la nueva se sintetiza en sentido $5' \rightarrow 3'$. La duplicación acaba cuando se encuentra una secuencia de nucleótidos denominada secuencia de terminación.



Aunque este es el procedimiento habitual, en algunos casos el ADN se sintetiza mediante un proceso denominado transcripción inversa, en la que se parte de ARN y se obtiene ADN gracias a la ADN-polimerasa-ARN-dirigida (transcriptasa inversa). Suele ocurrir en virus de ARN cancerígenos.

La transcripción consiste en copiar la secuencia de ADN en ARNm. A diferencia de lo que ocurre en la replicación, en donde se duplica la molécula de ADN completa, en la transcripción únicamente se copia un conjunto de genes determinado.

La copia que se realiza durante la transcripción se efectúa por complementariedad de bases y la unión de las mismas se realiza en unas zonas denominadas promotores. Se trata de secuencias comunes compuestas de 10 nucleótidos (secuencia -10 o caja TATA) o de 35 nucleótidos (secuencia -35 o caja de entrada). La síntesis del ARNm finaliza cuando se encuentra una secuencia de terminación. En este momento, el ARNm se separará y el ADN recuperará su forma de doble hélice original.

Finalizada la transcripción, el paso siguiente es la traducción o biosíntesis de proteínas. Durante este proceso, hay que tener en cuenta que no todas las secuencias de bases codifican proteínas, sino que hay segmentos que no tienen función aparente. Los primeros se denominan exones, mientras que los segundos son los intrones. En la síntesis de proteínas, se eliminan los intrones y en el ARNm se ensamblan únicamente los exones.

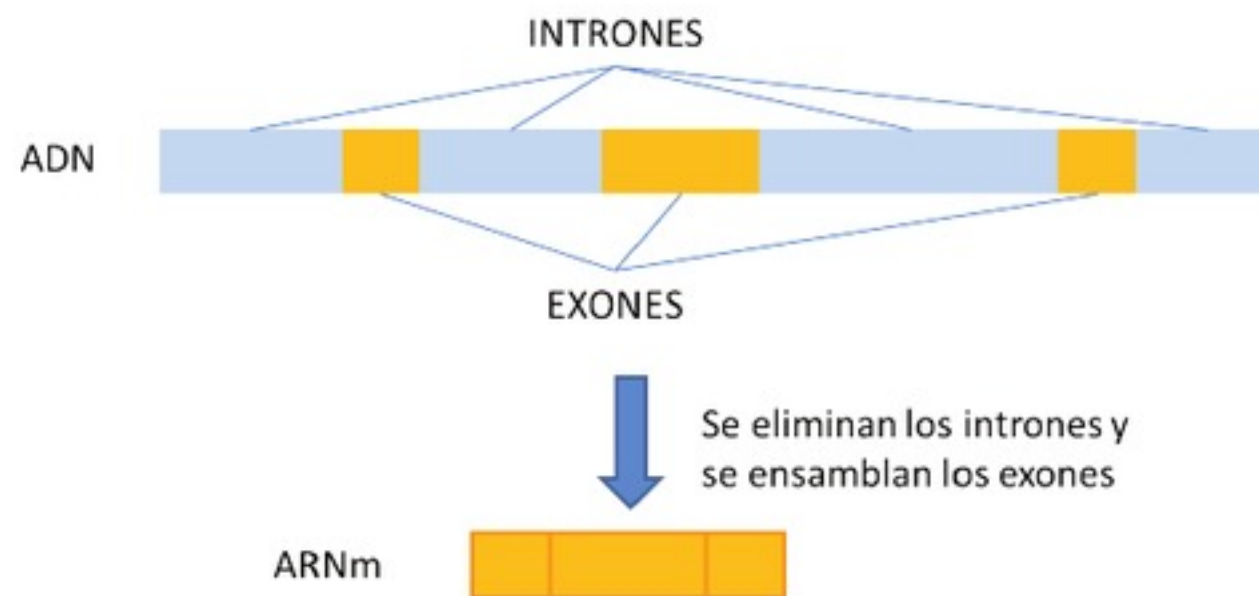


Figura 2.4. Síntesis de proteínas

2.5 REGULACIÓN GÉNICA

No todos los genes están activos en todo momento. Al contrario, de todos los genes que posee un organismo, solo unos pocos se expresan en cada instante. Este nivel de expresión se encuentra regulado debido al alto coste energético que supone la síntesis de proteínas. La regulación puede actuar durante la transcripción o durante la traducción.

En general, la regulación puede ser positiva o negativa. En la regulación positiva se aumenta la expresión de los genes, mientras que en la regulación negativa se disminuye. Los dos tipos de regulación, positiva y negativa, pueden darse simultáneamente en el mismo sistema celular. En ambos casos, los reguladores son proteínas que reconocen secuencias específicas de ADN.

Los mecanismos de regulación son diferentes en células procariotas y en células eucariotas. En el primer caso, el ARNm suele ser policistrónico (contiene varios genes), mientras que en el segundo cada secuencia de ARNm contiene la información de un solo gen (monocistrónico).



FORMATOS DE FICHEROS

Las plataformas de secuenciación generan grandes cantidades de información que exigen ser guardadas, parseadas y analizadas de manera eficiente. Una secuenciación típica producirá ficheros que ocuparán desde unos pocos *megabytes* a *terabytes* y contendrán miles o millones de lecturas, junto con información como identificadores de lectura, descripciones, anotaciones, metadatos, etc.

La Figura 3.1 muestra los principales formatos de ficheros en función de la etapa del análisis de la secuencia en que nos encontramos. Las plataformas de secuenciación, a partir de una muestra biológica, generan información en formato FASTA/FASTQ, principalmente. Estas secuencias biológicas (de nucleótidos o de aminoácidos) son susceptibles de sufrir un proceso de alineamiento (formatos SAM/BAM), análisis de variaciones (formato VCF) y anotación (formatos GFF/GFF3, GVF y BED).

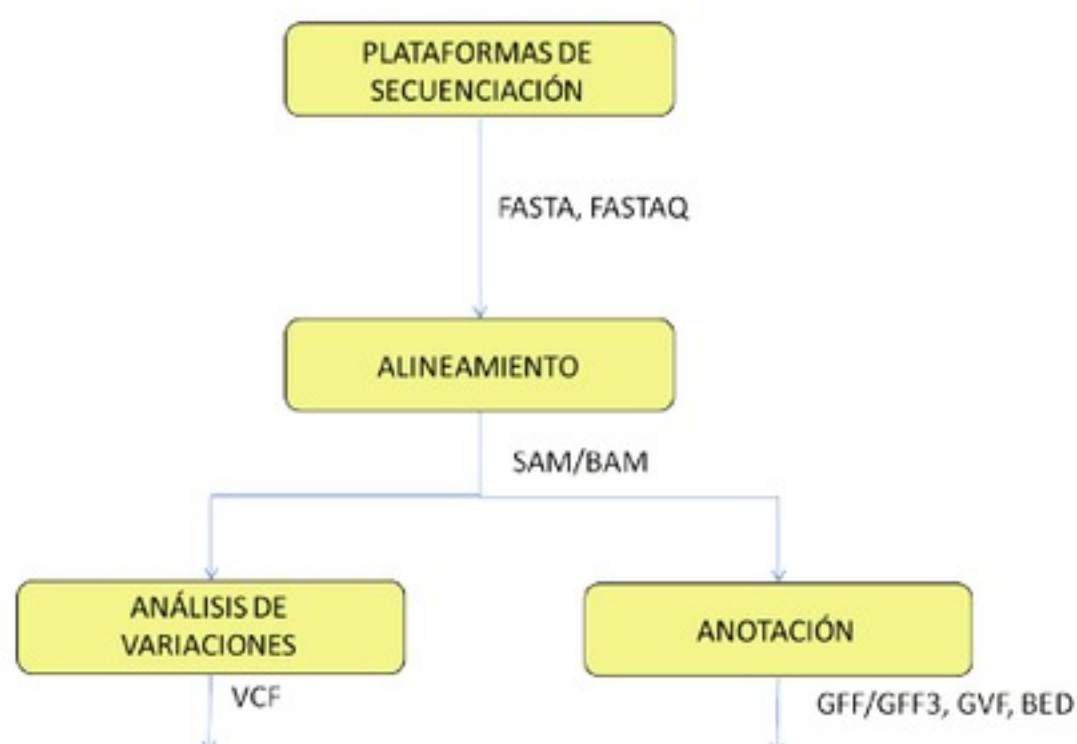


Figura 3.1. Formatos de ficheros en el análisis de datos genómicos



La lista de formatos que se estudiarán a lo largo de este libro no es exhaustiva, ni mucho menos. Si algo caracteriza la Bioinformática es la falta de un estándar global para representar la información, cualquiera que sea el estado de la misma. En su lugar, encontramos toda una miríada de formatos de ficheros. No obstante, hemos recogido aquí los más habituales.

3.1 DATOS EN BRUTO

El hecho de que una secuencia se pueda representar como una cadena de texto permite que, en realidad, todos estos ficheros con secuencias de nucleótidos y proteínas sean ficheros de texto plano. Estos ficheros únicamente pueden contener caracteres IUPAC (ver Tablas 3.1 y 3.2) y espacios y sirven para almacenar solo la secuencia:

```
ACAAGATGCCATTGTCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCC
CTGCCCCCTGGAGGGTACGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATA
AGGAAAAGCAGCCTCCTGACTTTCCTCGCTTGGTAGTGGACCTCCAGGCCAGTGCCGGGGCCCCT
CATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCATCCGCGCGCCGGG
ACAGAATGCCCTGCAGGAACCTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAA
```

Figura 3.2. Ejemplo de secuencia de nucleótidos

Este formato tiene algunas limitaciones obvias: no pueden incluirse ni el nombre de la secuencia (caracteres alfanuméricos) ni más de una secuencia por fichero. Por este motivo, y por algunos otros, se han ido incorporando otros formatos de fichero más complejos que permiten referir más información.

Código de la base	Significado
A	Adenina
C	Citosina
G	Guanina
T	Timina
U	Uracilo
R	G o A (purina)
Y	T o C (pirimidina)
K	G o T (cetona)
M	A o C (amino)

S	G o C (interacción fuerte)
W	A o T (interacción débil)
B	G, T o C (no A)
D	G, A o T (no C)
H	A, C o T (no G)
V	G, C o A (no T ni U)
N	A, G, C, T (cualquiera)
X	Máscara
-	Hueco

Tabla 3.1. Código IUB/IUPAC para nucleótidos

Código de aminoácido	Significado
A	Alanina
B	Asparagina
C	Cisteína
D	Ácido aspártico
E	Ácido glutámico
F	Fenilalanina
G	Glicina
H	Histidina
I	Isoleucina
K	Lisina
L	Leucina
M	Metionina
N	Asparagina
O	Pirrolisina
P	Prolina
Q	Glutamina
R	Arginina
S	Serina
T	Treonina
U	Selenocisteína
V	Valina
W	Triptófano
Y	Tirosina
Z	Glutamina
X	Cualquiera
*	Parada de traducción
-	Hueco

Tabla 3.2. Código IUB/IUPAC para aminoácidos

3.2 FASTA

Un fichero en formato FASTA es un fichero de texto que contiene largas cadenas de caracteres correspondientes a secuencias de nucleótidos o péptidos. Además, incluye una línea de definición en la que se encuentra el nombre de las secuencias y comentarios aclarativos sobre las mismas.

La línea de definición del formato FASTA se caracteriza por comenzar por el símbolo ‘>’ seguido de un nombre y una descripción. A continuación, se encuentra la secuencia escrita en texto plano:

```
>Human
ATGGCACATGCAGCGCAAGTAGGTCTACAAGACGCTACTTCCCCTATCATAGAAGAGCTTATCAC
CTTTCATGATCACGCCCTCATAATCATTTTCCTTATCTGCTTCCTAGTCCTGTATGCCCTTTTCC
TAACACTCACAACAAAATACTAATACTAACATCTCAGACGCTCAGGAAATAGAAACCGTCTGA
ACTATCCTGCCCCGCCATCATCCTAGTCCTCATCGCCCTCCCATCCCTACGCATCCTTTACATAAC
AGACGAGGTCAACGATCCCTCCCTTACCATCAAATCAATTGGCCACCAATGGTACTGAACCTACG
AGTACACCGACTACGGCGGACTAATCTTCAACTCCTACATACTTCCCCCATTATTCCTAGAACCA
GGCGACCTGCGACTCCTTGACGTTGACAATCGAGTAGTACTCCCGATTGAAGCCCCCATTCGTAT
AATAATTACATCACAAGACGTCTTGCACTCATGAGCTGTCCCCACATTAGGCTTAAAAACAGATG
CAATTCCCGGACGTCTAAACCAAACCACTTTCACCGCTACACGACCGGGGGTATACTACGGTCAA
TGCTCTGAAATCTGTGGAGCAAACCACAGTTTCATGCCCATCGTCCTAGAATTAATTCCCCTAA
```

Figura 3.3. Secuencia en formato FASTA

Se recomienda que todas las líneas de texto tengan menos de 80 caracteres para facilitar la visualización y la edición. Por otra parte, las secuencias deben estar escritas en el estándar IUB/IUPAC, con las siguientes excepciones:

- Se aceptan letras minúsculas, aunque luego se conviertan a mayúsculas.
- Empleo de un paréntesis o un guión para representar un hueco de longitud indeterminada.
- Se aceptan la U y el carácter ‘*’ en secuencias de aminoácidos.

Con este formato, es posible incluir varias secuencias en un mismo fichero, siempre y cuando tengan líneas de definición distintas.

3.3 FASTAQ

Se utiliza, sobre todo, para guardar las lecturas de las plataformas de secuenciación y se trata de una extensión del FASTA que incorpora, junto a la secuencia, una indicación de la calidad o PHRED de cada nucleótido de la secuencia. Tanto la letra de la secuencia como su calidad se codifican, por “simplicidad” y no “simplificidad”, en un carácter ASCII.

En un fichero FASTAQ, existen cuatro secciones, tal y como se muestra en el ejemplo (ver Figura 3.4). En la primera línea, se encuentra una arroba ('@') seguido de un identificador de la secuencia sin ningún límite de longitud, por lo que es posible incluir alguna anotación arbitraria o comentarios. A continuación, aparecen las líneas de la secuencia en formato FASTA y, para señalar el fin de las líneas de secuencia y el comienzo de los indicadores de calidad, la tercera sección empieza por un '+', opcionalmente, seguido del identificador de la secuencia y la descripción que ya se indicaron en la primera línea. Finalmente, vienen las líneas de calidad, que son tiras de caracteres ASCII imprimibles, en donde el carácter '!' es la calidad más baja, el '~' la más alta y el resto son los valores de calidad ordenados de izquierda a derecha por calidad creciente.

```
@secuencia_1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!' '*((( (**+)) %%%++) (%%%) .1***-+*' ')) **55CCF>>>>>CCCCCCC65
@secuencia_2
ATCGTAGTCTAGTCTATGCTAGTGCGATGCTAGTGCTAGTCGTATGCATGGCTATGTGTG
+
208DA8308AD8SF83FH0SD8F08APFIDJFN34JW830UDS8UFDSADPFIJ3N8DAA
```

Figura 3.4. Secuencia en formato FASTAQ

Es importante remarcar que el símbolo '@' podría aparecer en cualquier posición de la línea de calidad. En estos casos, no debe tratarse como marcador de una nueva secuencia y debe comprobarse que la longitud de la línea de calidad es idéntica a la longitud de la secuencia.



La longitud de la secuencia y la de la línea de calidad deben ser idénticas. En caso contrario, el fichero se puede considerar corrupto e inservible.

El indicador PHRED está relacionado, logarítmicamente, con la probabilidad de error en la secuenciación P :

$$Q = -10 \log_{10} P$$

$$P = 10^{-\frac{Q}{10}}$$

De esta manera, a partir de la línea de calidad es posible calcular la probabilidad de un error en la secuenciación. Por ejemplo, si escogemos los valores de calidad correspondientes a las tres primeras bases de la secuencia `secuencia_1`:

```
!''*(((((***+))%%#+))%%%).1***-+''')**55CCF>>>>>CCCCCCC65
```

Figura 3.5. Ejemplo de línea de calidad

Sus valores ASCII son, respectivamente, 33, 39 y 39. Teniendo en cuenta la relación anterior, las probabilidades de error en la secuenciación serán de 0,0501%, 0,0125% y 0,0125%.

Existen varias versiones de este formato, según el fabricante de la plataforma de secuenciación, pero la más utilizada es la Sanger, que acepta PHRED en el rango [0, 93] y caracteres ASCII entre el 33 y el 126. Por otro lado, Illumina 1.0 codifica el PHRED entre [-5, 62] y caracteres ASCII entre el 59 y el 126 mientras que Illumina 1.3+ codifica PHRED entre [0, 62] y caracteres ASCII entre 64 y 126.

3.4 SAM/BAM

El formato SAM (*Sequence Alignment/Map*) es un formato genérico utilizado para guardar alineamientos de secuencias de nucleótidos. Se trata de un formato cuyo parseo consume gran cantidad de recursos y es lento, por lo que se definió una versión binaria del mismo, el formato BAM.

Las principales ventajas de estos formatos son las siguientes:

- Flexibles, ya que permiten almacenar información de alineamiento generada por varios programas.
- Simples, lo que posibilita que gran cantidad del software existente los soporten.
- El tamaño del fichero es razonablemente reducido.
- No es necesario cargar todo el fichero para empezar a procesar la información que contiene.
- Es posible indexar el contenido por la posición genómica.

Un fichero SAM es un fichero delimitado por tabuladores que contiene una cabecera (opcional) y una sección con los datos del alineamiento propiamente dichos.

Si está presente, la cabecera es la primera de las secciones. Las líneas de la cabecera van todas precedidas por una arroba '@' y un código de dos letras. Todas las líneas de cabecera, excepto las que empiecen por @CO, deben estar tabuladas. A continuación, se encuentran pares Etiqueta:Valor.

Campo de la cabecera	Contenido	Significado
@HD	VN*	Versión
	SO	Ordenación de los alineamientos Valores aceptados: <i>unkown</i> (por defecto), <i>unsorted</i> , <i>queryname</i> y <i>coordinate</i>
@SQ		Diccionario de referencia de secuencias. El orden de las líneas @SQ determina el orden del alineamiento
	SN*	Nombre de referencia de la secuencia
	LN*	Longitud de la secuencia
	AS	Identificador de ensamblado del genoma
	M5	Control de errores codificado en MD5
	SP	Especies
	UR	URI de la secuencia
@RG		Grupo de lecturas
	ID*	Identificador del grupo de lecturas. Debe ser único a nivel de todos los grupos de lecturas
	CN	Nombre de la secuencia central que produce la lectura
	DS	Descripción
	DT	Fecha de ejecución
	FO	Orden del flujo
	KS	Bases de la secuencia
	LB	Librería
	PG	Programa usado para procesar el conjunto de lecturas
	PI	Predicción del tamaño medio de las inserciones
	PL	Plataforma de secuenciación
	PU	Unidad usada en la plataforma
	Sm	Muestra
@PG		Programa
	ID*	Identificador del programa de registro
	PN	Nombre del programa
	CL	Línea de comandos
	PP	Anterior @PG-ID. Debe coincidir con el ID de alguna etiqueta de encabezado @PG
	VN	Versión del programa
@CO		Comentario

Tabla 3.3. Estructura de la cabecera

Orden	Nombre	Descripción
1	QNAME	Nombre de la consulta
2	FLAG	Bandera de opciones
3	RNAME	Nombre de la secuencia
4	POS	Posición de la base más a la izquierda
5	MAPQ	Calidad del mapeo
6	CIGAR	Cadena CIGAR
7	RNEXT	Referencia al siguiente fragmento de la secuencia
8	PNEXT	Posición del siguiente fragmento
9	TLEN	Longitud de la plantilla
10	SEQ	Fragmento de la secuencia
11	QUAL	Calidad de la secuencia



En el ejemplo de la Figura 3.6 vemos que se trata de un fichero de la versión 1.0 (@HD VN:1.0) donde los alineamientos se han ordenado por el campo RNAME (SO:coordinate). Por otra parte, existen dos secuencias (@SQ SN:seq1 y @SQ SN:seq2) y se ha añadido el comentario “Ejemplo de fichero SAM”.

```
@HD VN:1.0 SO:coordinate
@SQ SN:seq1 LN:5000
@SQ SN:seq2 LN:5000
@CO Ejemplo de fichero SAM
B7_591:4:96:693:509 73 seq1 1 99 36M * 0
0 CACTAGTGGCTCATTGTAAATGTGTGGTTTAAC TCG <<<<<<<<<<<<<<<<<<<<<<<<<
<<<<5<<<<<<;:<<;7 MF:i:18 Aq:i:73 NM:i:0 UQ:i:0 H0:i:1
```



```
H1:i:0
EAS54_65:7:152:368:113 73 seq1 3 99 35M * 0
0 CTAGTGGCTCATTGTAAATGTGTGGTTTAACTCGT <<<<<<<<<0<<<<655<<
<<<:9<<3/:<6): MF:i:18 Aq:i:66 NM:i:0 UQ:i:0 H0:i:1
H1:i:0
```

Figura 3.6. Fragmento de un fichero en formato SAM

En cuanto a la sección de alineamientos, podemos deducir la información resumida en la Tabla 3.5:

Campo	Alineamiento 1	Alineamiento 2
QNAME	B7_591:4:96:693:509	EAS54_65:7:152:368:113
FLAG	73	73
RNAME	seq1	Seq1
POS	1	3
MAPQ	99	99
CIGAR	36M	35M
MRNM/RNEXT	*	*
MPOS/PNEXT	0	0
ISIZE/TLEN	0	0
SEQ	CACTAGTGGCTCATTGTA AATGTGTGGTTTAACTCG	CTAGTGGCTCATTGTAAA TGTGTGGTTTAACTCGT
QUAL	<<<<<<<<<<<<<<<<;<<<<<<<< 5<<<<<<;:<7	<<<<<<<<<<0<<<<655<<7<<< :9<<3/:<6):
TAGs	MF:i:18 Aq:i:73 NM:i:0 UQ:i:0 H0:i:1 H1:i:0	MF:i:18 Aq:i:66 NM:i:0 UQ:i:0 H0:i:1 H1:i:0

Tabla 3.5. Contenido de la sección de alineamientos

Uno de los puntos en los que merece la pena detenerse, por su complejidad, es en la obtención de la cadena CIGAR (*Compact Idiosyncratic Gapped Alignment Report*), que describe cómo se mapea una secuencia respecto a un genoma de referencia.

La especificación establece que para calcular la cadena CIGAR de un alineamiento hay que utilizar la Tabla 3.6:

Operador	BAM	Descripción
M	0	Coincidencia en el alineamiento
I	1	Inserción
D	2	Borrado
N	3	Salto de una región que está presente en la referencia pero no en la secuencia que se alinea
S	4	<i>Soft clipping</i>
H	5	<i>Hard clipping</i>
P	6	Relleno de la referencia
=	7	Coincidencia entre la referencia y la secuencia que se alinea
X	8	No coincidencia entre la referencia y la secuencia que se alinea

Tabla 3.6. Operadores de cálculo de la cadena CIGAR

Supongamos que un secuenciador nos proporciona la siguiente lectura:

ACTAGAATGGCT

Figura 3.7. Ejemplo de muestra de un secuenciador

Y que queremos alinearla con la siguiente secuencia de referencia:

CCATACTGAACTGACTAAC

Figura 3.8. Secuencia con la que se quiere alinear la salida del secuenciador

Para calcular la cadena CIGAR correspondiente a este alineamiento, resulta útil representarla en forma de tabla. En la fila superior se muestran las posiciones (*P*); en la segunda fila, las bases de la secuencia de referencia y, en la tercera, la secuencia que se alinea.

P	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
R	C	C	A	T	A	C	T	G	A	A	C	T	G	A	C	T	A	A	C
L	A	C	T	A	G	A	A	T	G	G	C	T							

Figura 3.9. Cálculo de la cadena CIGAR

Tomando como referencia la posición 0, la cadena CIGAR sería 1X1=8X2=7D, es decir, que hay una base que no coincide en las dos secuencias, otra base que sí coincide, seguida de 8 bases que no coinciden, 2 bases que sí coinciden y 7 borrados (bases que están en la referencia pero no en la secuencia que se alinea).

3.5 GFF/GFF3

Un fichero en formato GFF (*General Feature Format*) está formado por líneas con nueve campos cada una, todos ellos obligatorios. Se trata de un formato muy restrictivo puesto que si, por ejemplo, los campos se separan por espacios en lugar de por tabuladores, algunas herramientas no podrán procesarlo adecuadamente.

La Figura 3.10 muestra un ejemplo de fichero en formato GFF:

```
SEQ1 EMBL atg 103 105 . + 0
SEQ1 EMBL exon 103 172 . + 0
SEQ1 EMBL splice5 172 173 . + .
SEQ1 netgene splice5 172 173 0.94 + .
SEQ1 genie sp5-20 163 182 2.3 + .
SEQ1 genie sp5-10 168 177 2.1 + .
SEQ2 grail ATG 17 19 2.1 - 0
```

Figura 3.10. Fichero en formato GFF

El principal problema del formato GFF es que, a pesar de encontrarse muy extendido, existen algunas variantes incompatibles entre sí. Para resolver este inconveniente se creó un nuevo formato, el GFF3, caracterizado principalmente por:

- Proporciona un mecanismo para representar más de un nivel de agrupamiento jerárquico de características y características de características.
- Separa las ideas de miembro de un grupo y de nombre de característica.
- Restringe el tipo de características a un vocabulario controlado.
- Permite que una característica dada, como un exón, pertenezca a más de un grupo a la vez.
- Proporciona una convención explícita para los alineamientos de pares.
- Proporciona una convención explícita para características que ocupan regiones disjuntas.

La Tabla 3.7 resume brevemente el significado de estos campos:

Nombre del campo	Descripción
Seqname	Nombre de la secuencia
Source	Programa que ha generado la secuencia
Feature	Tipo de secuencia (gen, CDS, etc.)
Start	Posición de inicio de la secuencia
End	Posición final de la secuencia
Score	Calificación de la secuencia. Cuando no existe ninguna calificación, se especifica un '.' o un 0
Strand	Dirección de la secuencia. Los valores válidos son '+', '-', o '.' (si no se sabe o no importa).
Frame	Toma los valores '0', '1', '2' o '.' '0': indica que la primera base de la región especificada se corresponde con la primera base de un codón '1': significa que hay una base adicional, es decir, que la segunda base de la región especificada se corresponde con la primera base de un codón '2': quiere decir que la tercera base de la región especificada es el primera base de un codón
Attributes	Opcional. Otros atributos definidos por el usuario
Comments	Opcional. Cualquier comentario acerca de la secuencia

Tabla 3.7. Campos de un fichero en formato GFF3

Teniendo en cuenta la Tabla 3.7, la primera línea del fichero de ejemplo representa una secuencia llamada SEQ1 obtenida de la base de datos de secuencias de EMBL cuya característica atg, comprendida entre los nucleótidos 103 y 105, sin calificación, pertenece a la dirección '+'. Al tener el campo *frame* el valor cero, se trata de un exón.



Utilizando la información de la Tabla 3.7, interpreta el resto de líneas del fichero. Si tienes alguna duda, puedes consultar directamente la especificación oficial del formato en <https://www.sanger.ac.uk/resources/software/gff/spec.html>.

3.6 GVF

GVF es un tipo de formato GFF3 que incluye pragmas y atributos adicionales (ver Tabla 3.8). De hecho, un fichero GVF tiene la misma estructura de líneas delimitadas por tabuladores con nueve campos y todas las restricciones que se aplican a un fichero GFF3, también lo hacen a un fichero GVF.

La cabecera de un fichero GVF está compuesta de pragmas precedidas por una doble almohadilla. Contienen metadatos y la única obligatoria es `##gvf-version 1.07`, que indica la versión de la especificación del formato del fichero que se ha utilizado.



Es muy importante remarcar el hecho de que la almohadilla debe ser doble o, de lo contrario, la línea será ignorada, incluso si contiene una única almohadilla.

A la cabecera le siguen las líneas de características que contienen la información sobre las variaciones:

- *Seqid*: cromosoma o contig en el que se ha encontrado la variación.
- *Source*: fuente de la que procede la variación.
- *Type*: tipo de variación.
- *Start*: comienzo de la variación.
- *End*: fin de la variación.
- *Score*: probabilidad de error en la variación, calculada en escala PHRED.
- *Atributos*: conjunto de pares etiqueta=valor según las directrices recogidas en la Tabla 3.8 (los campos obligatorios están marcados con un asterisco):

Etiqueta	Descripción
ID*	Identificador único a nivel de fichero
Variant_seq	Secuencia de la variación. Si tiene más de 50 bases, se puede resumir con un '~'. Por otra parte, si se trata de un borrado respecto de la secuencia de referencia, el campo toma el valor '-'
Reference_seq	Secuencia de referencia
Variant_reads	Número de lecturas que soporta cada variación
Total_reads	Número total de lecturas
Genotype	Genotipo de la variación
Variant_freq	Frecuencia de la variación en una población de individuos
Variant_effect	Consecuencias de la variación en la secuencia de referencia
Variant_copy_number	Para las regiones en las que existen varias copias de la variación del genoma, este campo indica el número de copia
Reference_copy_number	Para las regiones en las que existen varias copias de la variación del genoma, este campo indica el número de copia en forma de identificador
Nomenclature	Indica la denominación HGVS de la variación

Tabla 3.8. Atributos de una variación



La especificación completa del formato GVF está accesible en <http://www.sequenceontology.org/resources/gvf.html>.

La Figura 3.11 muestra un ejemplo de variaciones descritas en formato GVF, según la especificación 1.07 del formato del fichero. Si analizamos la primera línea de características, obtendremos que se trata de una variación en el cromosoma 16, generada con *samtools*, de tipo SNV, que comienza en la posición 49291141 y acaba en la 49291141, con un indicador PHRED '+' (probabilidad de error del 0,005%), caracterizada porque se encuentran dos bases (A y G) en lugar de una G.

```
##gvf-version 1.07
##genome-build NCBI B36.3
##sequence-region chr16 1 88827254
chr16 samtools SNV 49291141 49291141 . + . ID=ID_1;Variant
seq=A,G;Reference_seq=G;
chr16 samtools SNV 49291360 49291360 . + . ID=ID_2;Variant
seq=G;Reference_seq=C;
chr16 samtools SNV 49302125 49302125 . + . ID=ID_3;Variant
seq=T,C;Reference_seq=C;
```



```
chr16 samtools SNV 49302365 49302365 . + . ID=ID_4;Variant
seq=G,C;Reference_seq=C;
chr16 samtools SNV 49302700 49302700 . + . ID=ID_5;Variant
seq=T;Reference_seq=C;
chr16 samtools SNV 49303084 49303084 . + . ID=ID_6;Variant
seq=G,T;Reference_seq=T;
chr16 samtools SNV 49303156 49303156 . + . ID=ID_7;Variant
seq=T,C;Reference_seq=C;
chr16 samtools SNV 49303427 49303427 . + . ID=ID_8;Variant
seq=T,C;Reference_seq=C;
chr16 samtools SNV 49303596 49303596 . + . ID=ID_9;Variant
seq=T,C;Reference_seq=C;
```

Figura 3.11. Ejemplo de fichero GVF



HGVS es una especificación de variaciones muy compleja cuya descripción queda fuera del ámbito de este libro. Remitimos al lector al sitio web <http://www.hgvs.org/mutnomen/>.

3.7 VCF

El formato VCF (*Variant Call Format*) se emplea para recoger información de variaciones. Como en los casos anteriores, se trata de un fichero de texto dividido en líneas cuyos campos se separan por tabuladores.

Los campos empleados por el formato VCF se resumen en la Tabla 3.9:

Columna	Campo	Descripción
1	CHROM	Nombre del cromosoma
2	POS	Posición, más a la izquierda, de la variación
3	ID	Identificador de la variación
4	REF	Alelo de referencia
5	ALT	Alelos alternativos, separados por comas
6	QUAL	Indicador de la calidad de la variación
7	FILTER	Filtros aplicados
8	INFO	Información sobre la variación
9	FORMAT	Formato de los campos del genotipo (opcional)
10	SAMPLE	Ejemplos de genotipos e información por muestra

Tabla 3.9. Campos de un fichero VCF

Por otra parte, las etiquetas empleadas en el campo INFO son las de la Tabla 3.10:

Etiqueta	Descripción
AA	Alelo ancestral
AC	Número de alelos en el genotipo
AF	Frecuencia del alelo
AN	Número total de alelos en el genotipo
BQ	Calidad RMS de esta posición
CIGAR	Cadena CIGAR que describe cómo alinear el alelo con el alelo de referencia
DB	Entrada dbSNP
END	Posición final de la variación
H2	Identificador HPMAP2
H3	Indentificador HPMAP3
MQ	Calidad RMS
MQ0	Número MAPQ
NS	Número de muestras con datos
SB	Hebra en esta posición
SOMATIC	Indica que el registro es una mutación somática para los estudios relacionados con el cáncer
VALIDATED	Variación contrastada experimentalmente
1000G	Miembro de 1000 Genomes

Tabla 3.10. Formato del campo INFO



El proyecto HapMap (<http://hapmap.ncbi.nlm.nih.gov/>) es una iniciativa internacional cuyo objetivo es recopilar información sobre genes y enfermedades humanas.

La Figura 3.12 muestra un ejemplo de fichero VCF sencillo.

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
```



```
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of
Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral
Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership,
build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2
membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have
data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype
Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype
Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT
NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2
GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017
GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0
667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T
GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTCT G,GTACT 50 PASS NS=3;DP=9;AA=G
GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

Figura 3.12. Ejemplo de fichero VCF

Los ficheros VCF suelen comprimirse e ir acompañados de un fichero de índices en tabix, lo que permite que un navegador genómico pueda interpretar únicamente los fragmentos de la información que se están viendo en ese momento.



La especificación completa del formato VCF está accesible en <http://www.1000genomes.org>.

3.8 BED

El formato BED proporciona una manera flexible de describir las anotaciones sobre variaciones.

Cada línea de un fichero BED tiene tres campos obligatorios y nueve campos opcionales y están resumidos en la Tabla 3.11, donde los campos obligatorios están marcados con un asterisco.

Campo	Descripción
Chrom*	Nombre del cromosoma o esqueleto
chromStart*	Inicio de la variación dentro del cromosoma o esqueleto
chromEnd*	Fin de la variación dentro del cromosoma o esqueleto
Name	Nombre de la línea del fichero
Score	Puntuación entre 0 y 100. Se utiliza en la representación gráfica del fichero
Strand	Orientación de la hebra ('+' o '-')
thickStart	Inicio en que se comienza a dibujar la lectura
thickEnd	Final del dibujo de la lectura
itemRgb	Color que tendrá la lectura cuando se visualice (solo si se ha indicado que itemRgb="On" en la cabecera)
blockCount	Número de exones presentes en la línea del fichero
blockSizes	Lista, separada por comas, de los tamaños de los exones. El número de elementos de la lista debe ser <i>blockCount</i>
blockStarts	Lista, separada por comas, de las posiciones de inicio de cada exón, relativas a <i>chromStart</i> . El número de elementos de la lista debe ser <i>blockCount</i>

Tabla 3.11. Campos de un fichero BED



La especificación completa del formato BED está accesible en <http://genome.ucsc.edu/FAQ/FAQformat.html#format1>.

La Figura 3.13 muestra un ejemplo de fichero BED:

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration"
visibility=2 itemRgb="On"
chr7 127471196 127472363 Pos1 0 + 127471196 127472363 255,0,0
chr7 127472363 127473530 Pos2 0 + 127472363 127473530 255,0,0
chr7 127473530 127474697 Pos3 0 + 127473530 127474697 255,0,0
chr7 127474697 127475864 Pos4 0 + 127474697 127475864 255,0,0
chr7 127475864 127477031 Neg1 0 - 127475864 127477031 0,0,255
chr7 127477031 127478198 Neg2 0 - 127477031 127478198 0,0,255
chr7 127478198 127479365 Neg3 0 - 127478198 127479365 0,0,255
chr7 127479365 127480532 Pos5 0 + 127479365 127480532 255,0,0
chr7 127480532 127481699 Neg4 0 - 127480532 127481699 0,0,255
```

Figura 3.13. Ejemplo de fichero BED

La representación gráfica de este fichero en el UCSC Genome Browser es la de la Figura 3.14. Como puede observarse, los colores se corresponden con los indicados en las diferentes líneas del fichero.

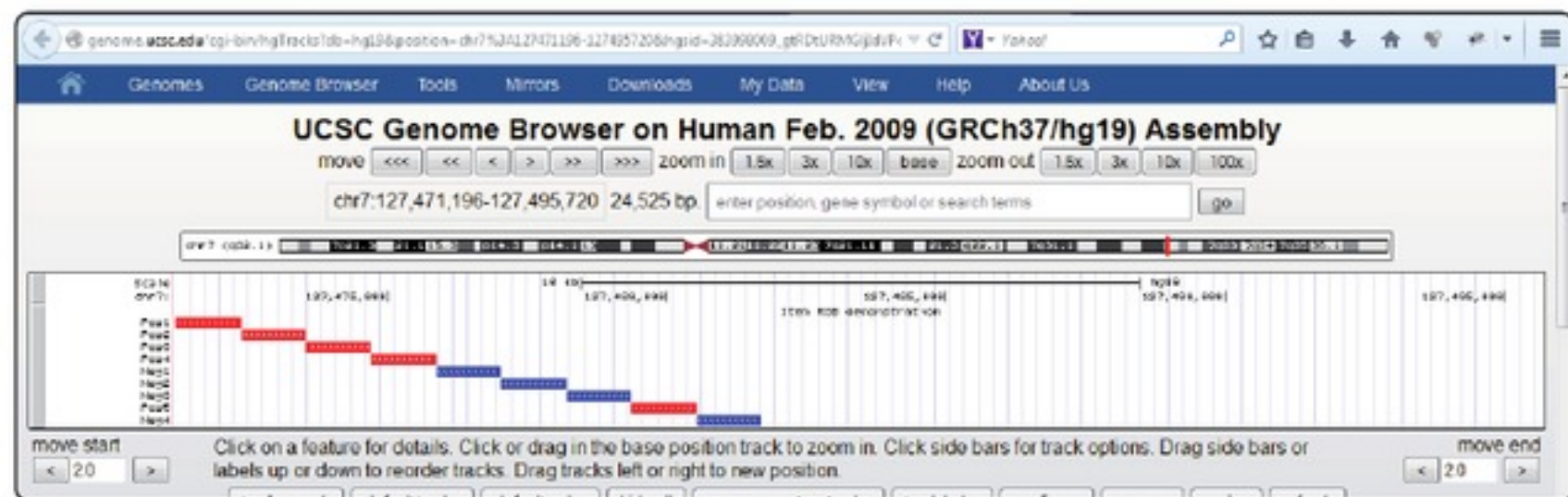


Figura 3.14. Representación gráfica del fichero BED anterior



BASES DE DATOS GENÓMICAS

El primer genoma se secuenció en 1995 en el Institute of Genomic Research y correspondía a la bacteria *Haemophilus influenzae*. Solo unos pocos meses más tarde, se completó el genoma de la levadura *Saccharomyces cerevisiae*. Desde entonces la cantidad de genomas secuenciados ha crecido a pasos agigantados, incluyendo la secuenciación completa del genoma humano en diciembre de 2013. Tanto es así que en apenas 20 años se ha producido una verdadera explosión de la información genómica gracias, entre otras cosas, al menor coste de la tecnología necesaria.

Merece la pena detenernos en observar algunos ejemplos que nos darán idea de la magnitud de la información a que nos referimos así como del crecimiento vertiginoso a que está sometido dicho volumen de información.

Una de las bases de datos genómicas más importantes, y que estudiaremos con detalle en un capítulo posterior, es la NCBI-GenBank. En la Figura 4.1 se observa que tanto el número de bases como el número de secuencias almacenadas ha evolucionado exponencialmente desde la versión 3 liberada en 1982. Junto con GenBank, se muestra también la evolución de los registros del proyecto WGS, en donde se maneja información de genomas procariotas y eucariotas y que comenzó en abril de 2002.

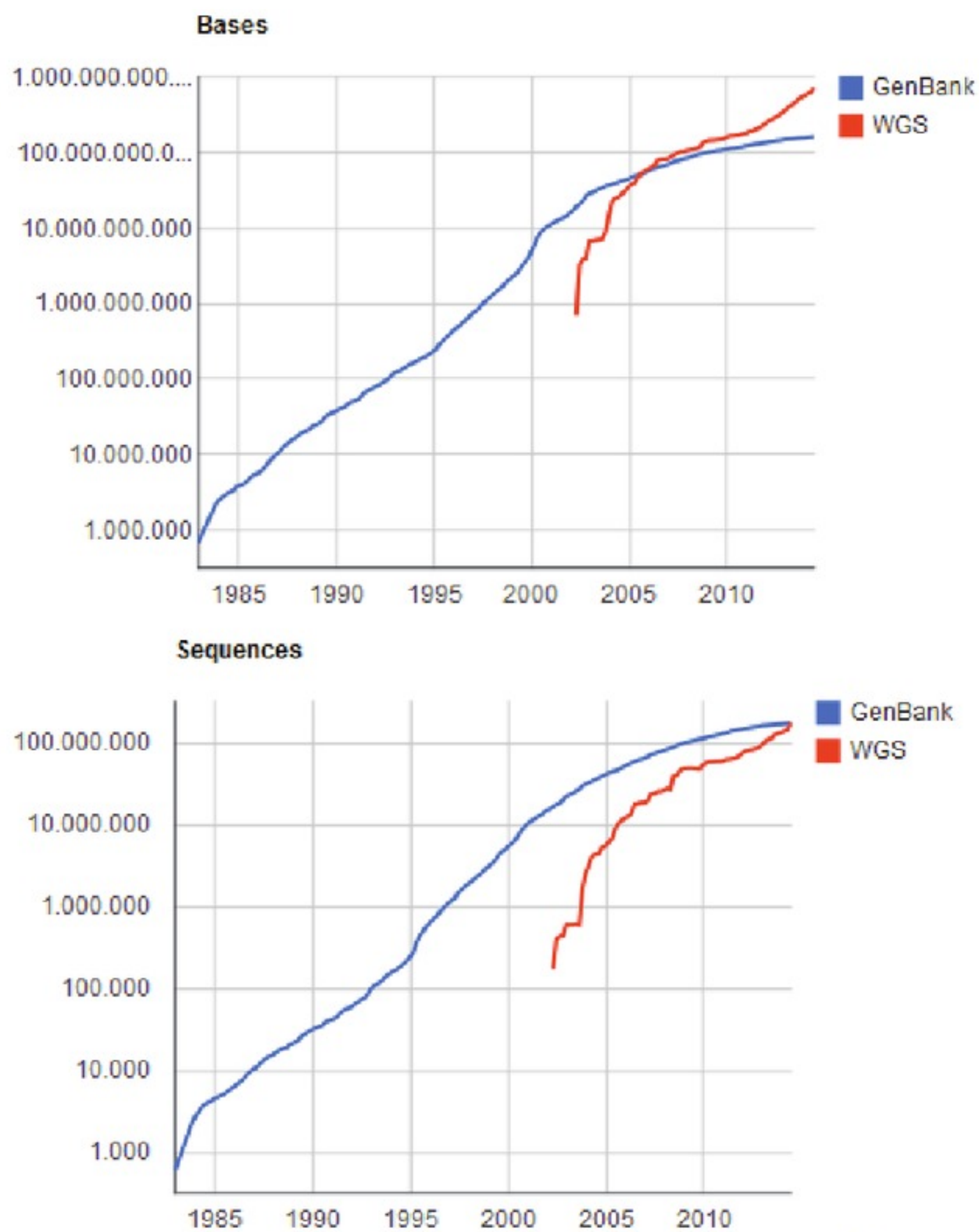


Figura 4.1. Evolución de la información almacenada en GenBank y WGS

Por otra parte, la Figura 4.2 resume los genomas registrados en la base de datos Entrez y el sistema de búsqueda de secuencias por excelencia, que también está alojado en el NCBI:

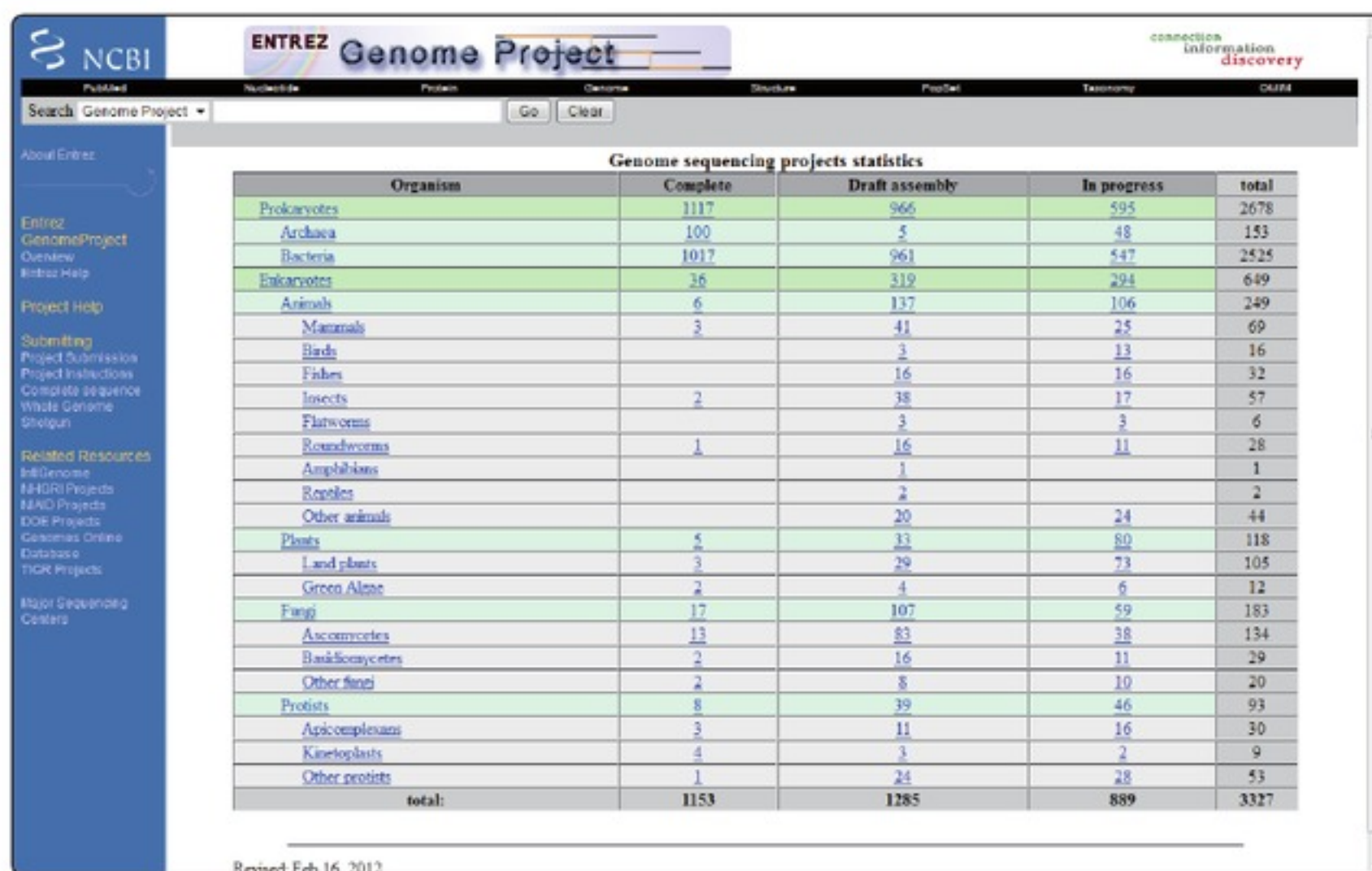


Figura 4.2. Estadísticas de la base de datos de genomas del NCBI

Esta exposición de información que, en cierta manera, podría verse como una oportunidad para desarrollar un conocimiento más profundo de distintos organismos, también constituye un riesgo de encontrarse información dispersa, desactualizada y compleja de procesar.

A medida que la cantidad de los datos aumenta, la tarea de identificar las relaciones críticas entre ellos se vuelve compleja. Organizar la información biológica en bases de datos dedicadas en las que se almacena información relacionada ha sido muy útil. Sin embargo, el número de estas bases de datos ha alcanzado una cifra considerable (del orden de miles de ellas al año), lo que plantea otros retos como la extracción de información útil o la minería de datos eficiente.

En gran medida, las dificultades actuales del análisis de datos biológicos surgen, simplemente, de la necesidad de cruzar la información presente en distintas bases de datos que, por si fuera poco, se encuentran dispersas y, además, con más frecuencia de la necesaria ofrecen datos incoherentes e incluso contradictorios.

La solución a estos problemas de dispersión de recursos y variedad de formatos de la información son las bases de datos genómicas, cuyas características, variedad y principios básicos de diseño abordaremos a lo largo de este capítulo.

4.1 ¿QUÉ ES UNA BASE DE DATOS GENÓMICA?

Una base de datos genómica es un repositorio de información que incluye todos o parte de los datos de una secuencia de ADN de uno o más organismos. Generalmente, en una base de datos genómica también puede encontrarse información adicional (anotaciones) que o bien describen características de la propia secuencia de ADN o bien propiedades biológicas de los organismos a los que pertenece dicha secuencia. Las anotaciones genómicas pueden clasificarse en dos grandes grupos:

- Anotaciones cromosómicas locales, que están asociadas a una determinada región de un cromosoma. La ubicación de los genes o anotaciones sobre la estructura del gen que indiquen las fronteras exón/intrón constituyen algunos ejemplos de este tipo de anotaciones.
- Anotaciones no directamente asociadas con una región genómica como datos sobre la estructura de la proteína o sobre vías metabólicas.

Algunas bases de datos genómicas proporcionan una interfaz de usuario que permite realizar consultas, visualizar las anotaciones, etc. Esta interfaz recibe el nombre de navegador genómico.



Una base de datos genómica es un repositorio de información que incluye todos o parte de los datos de una secuencia de ADN de uno o más organismos.

4.2 CLASIFICACIÓN DE LAS BASES DE DATOS GENÓMICAS

Las primeras bases de datos genómicas que aparecieron se centraban en la integración de los datos de una única especie. Ejemplos representativos son WormBase (dedicada al gusano *Caenorhabditis elegans*) o FlyBase (especializada en la mosca de la fruta, la *Drosophila melanogaster*). Sin embargo, la aparición de otras

bases de datos y la evidencia de la relación entre los genomas de distintas especies han puesto de manifiesto que para la comprensión del genoma de una especie es vital compararlo con la evolución de sus parientes cercanos. Así, podemos encontrar bases de datos genómicas globales, que contienen datos de varios organismos y diferentes tipos de secuencia y, bases de datos genómicas especializadas en determinados organismos, categorías o funciones específicas de secuencias o datos generados por tecnologías de secuenciación concretas. La Tabla 4.1 recoge algunos de los ejemplos más representativos.

Globales	Nucleótidos	GenBank EMBL: European Molecular Biology Laboratory DDBJ: DNA Data Bank of Japan
	Proteínas	Entrez Protein Swiss-Prot PIR-International UniProt
	Estructura de proteínas	MMDB: Molecular Modeling Database (basada en PDB) PDB: Protein Data Bank
	Genomas y mapas	Entrez Genome
	Bibliografía	PUBMED
	Rutas metabólicas	KEGG
	Enfermedades genéticas humanas	OMIM
Especializadas	Organismos	Human Genome Sequencing GDB: Genome Database (human mapping information) MGD: Mouse Genome Database SGD: Saccharomyces Genome Database
	Categorías o funciones de secuencias	TRANSFAC: Transcription Factors Vector Database
	Datos generados por tecnologías de secuenciación concretas	EST: Expressed Sequence Tags GSS: Genome Survey Sequences STS: Sequence Tagged Sites HTG: High Throughput Sequence

Tabla 4.1. Ejemplos de bases de datos genómicas

Finalmente, otra manera muy extendida de clasificar las bases de datos genómicas es atendiendo al origen de los datos que almacenan. Según este criterio distinguimos entre bases de datos primarias y bases de datos secundarias.

Las bases primarias almacenan datos en bruto procedentes de experimentos en laboratorios y se dividen en bases de datos de secuencias (como Swiss-Prot) y bases de datos estructurales (como Protein Databank para la estructura de proteínas). Generalmente, contienen enormes cantidades de información y se actualizan con una elevada frecuencia.

Por su parte, las bases de datos secundarias contienen información derivada de una o más bases de datos primarias que se obtiene como resultado del procesamiento de estas últimas. A diferencia de las bases de datos primarias, su frecuencia de actualización es más reducida. Su mayor ventaja es que, al contener información tratada, ahorran a los científicos tiempo y esfuerzo en el análisis de los datos.

4.3 CARACTERÍSTICAS DE LA INFORMACIÓN GENÓMICA

En teoría, la construcción de una base de datos biológica no debería ser muy distinta de la construcción de una base de datos para cualquier otro dominio de aplicación como un sistema de información bancario o el de una agencia de seguros. Desgraciadamente, nada más lejos de la realidad, ya que la información biológica presenta unas características únicas que dificultan, y mucho, la implementación de este tipo de sistemas de información.

Uno de los aspectos que más merece la pena destacar es el grado de incertidumbre en las reglas de negocio, a diferencia de lo que ocurre en otros entornos. Las reglas de negocio son las restricciones que hay que aplicar sobre los datos y que se deducen de las entidades del mundo real que representan dichos datos, así como las debidas a las necesidades específicas de la organización que explotará la base de datos. En un contexto biológico, las reglas de negocio que marcan las restricciones sobre los datos presentan cierto nivel de incertidumbre, debido, fundamentalmente, a las siguientes razones:

- La interpretación de la información almacenada puede cambiar con el tiempo gracias a un descubrimiento científico que altera las relaciones entre los datos. Por ejemplo, algunos genes estructurales de ARN no se almacenaban en las bases de datos hasta hace relativamente poco tiempo, cuando se descubrieron otros genes de ARN con funciones muy diversas. Esta ampliación del conocimiento tiene un claro impacto en el modelado del sistema.
- La información almacenada se considera invariable basándose en unos ciertos resultados experimentales, ya que hay medidas más fiables

que otras. Sobre esta información también se genera otro resultado de la interpretación de la primera. Si un experimento demostrara que la supuesta invariabilidad era errónea, la interpretación de la misma exigiría una revisión que podría tener un impacto considerable en el modelo.



Una manera de reducir la incertidumbre es almacenarlos junto a información de manera desestructurada en forma de comentarios de texto. Para evitar que esta desestructuración afecte a las consultas complejas, es preferible categorizarla de alguna manera. En la Práctica 1 se mostrará cómo.

Por otra parte, el elevado nivel de complejidad de los datos biológicos comparado con otros dominios de aplicación de las bases de datos supone un reto para el modelado de las estructuras de datos y sus relaciones que ha sido abordado desde diferentes perspectivas (ficheros planos, bases de datos relacionales, bases de datos orientadas a objetos, etc.). Un modelado poco preciso puede llevar a una pérdida de información que desemboque en un fallo total del diseño.

En general, distintas bases de datos emplean diferentes formatos para almacenar y representar la información (este punto lo trataremos en un capítulo posterior). Por tanto, los tipos de datos escogidos en el diseño de la base de datos deben ser suficientemente flexibles. De hecho, en ocasiones es necesario utilizar distintos tipos de datos para la misma información dependiendo del contexto de uso, por lo que resulta habitual construir varios esquemas conceptuales de un mismo esquema físico.

Otro obstáculo es que los usuarios de las bases de datos genómicas tienen, normalmente, un conocimiento reducido del diseño del esquema, lo que obliga a que la interfaz de acceso a la base de datos (generalmente, a través de la web) sea flexible, intuitiva y maneje una gran variedad de consultas potencialmente complejas así como que permita realizar comparaciones con los resultados obtenidos de otras bases de datos.

Finalmente, resulta bastante común realizar cambios en el esquema y someterlo a un proceso continuo y recursivo de rediseño. La principal consecuencia es que es recomendable, por no decir obligatorio, seguir una política de control de versiones con publicación de las actualizaciones respecto de versiones anteriores.



Las principales características de la información almacenada en las bases de datos genómicas son las siguientes:

- Muy compleja y, en ocasiones, redundante.
- Variabilidad de la información genómica.
- Elevada tasa de cambio que puede implicar cambios en el esquema de la base de datos.
- Conocimiento limitado del esquema por parte de los usuarios de la base de datos.
- Importancia de consultas complejas.

4.4 CONSTRUCCIÓN DE UNA BASE DE DATOS GENÓMICA

La construcción de una base de datos genómica es una tarea compleja en la que pueden distinguirse algunas subtarear comunes a la gran mayoría de bases de datos genómicas, y que son:

- Diseño e implementación de la arquitectura de almacenamiento que albergará los datos.
- Mantenimiento y actualización de la base de datos a medida que se publique información adicional.

En muchos casos, la responsabilidad de estas tareas recae en equipos de proyectos distintos. Esta complejidad de tareas y dispersión de responsabilidades tiene como principal consecuencia que no todas las bases de datos genómicas sean igual de fiables. De hecho, cabe distinguir entre archivos genómicos y bases de datos maduras.

Los archivos genómicos son bases de datos genómicas que constituyen un repositorio de información en ocasiones redundante (por ejemplo, porque contenga varias secuencias del mismo gen, una de un laboratorio diferente), no se mantiene ningún control sobre los registros que se guardan, no existe un vocabulario controlado, etc. No se trata, ni mucho menos, de bases de datos con información errónea, pero sí conviene tener en cuenta estos aspectos.

Por su parte, las bases de datos maduras están sometidas a procedimientos de control rigurosos que consiguen eliminar la redundancia de la información y garantizar la coherencia de los datos guardados. Además, dicha información es revisada por equipos de expertos.

4.5 MODELADO DE INFORMACIÓN GENÓMICA

Un modelo adecuado es crítico para la construcción y mantenimiento de la base de datos genómica. Obviamente, para realizar un buen modelado resulta imprescindible un mínimo conocimiento del dominio del problema, es decir, del contexto biológico en que se enmarca la información genómica. Por ello, remitimos al lector a la primera parte de este libro, en la que encontrará una explicación detallada de los conceptos biológicos básicos.

A la hora de elegir la técnica de modelado, conviene centrarse en los aspectos de los datos o del análisis de los mismos relevantes para la aplicación y enlazar a otras bases de datos biológicas para el resto. Sin embargo, esta limitación del enfoque no debe confundirse con un diseño de la base de datos que cubra los requisitos más inmediatos de la aplicación. Al contrario, el diseño de la base de datos puede ignorar o simplificar la información que quede fuera del ámbito del proyecto, pero debe establecer una representación completa de la parte de información biológica que representa, incluso aunque dicha aplicación no vaya a utilizar parte de la información almacenada. La consecuencia principal es que el diseño es mucho más robusto y flexible.

En cuanto a las técnicas de modelado, destacaremos las siguientes:

- **Modelado Entidad-Relación:** es una de las técnicas de modelado más extendidas en la Ingeniería del Software. Resulta muy adecuada cuando existen entidades bien definidas con relaciones relativamente simples entre ellas. Como hemos visto en el punto anterior, este no es el caso, por lo que es necesario extender el modelo Entidad-Relación. De hecho, en la Práctica 1 utilizaremos el modelo Entidad-Relación Extendido, ampliamente documentado en la Bibliografía.
- **Modelado UML:** UML (*Unified Modelling Language*) es un lenguaje de modelado que captura información de un sistema bajo estudio desde diversas perspectivas y con distintos niveles de detalle. Por otra parte, ofrece una representación ideal para perspectivas orientadas a objetos.
- **Modelado XML (*eXtensible Markup Language*):** XML es un lenguaje de marcado que se emplea para representar la estructura de la información que maneja el sistema bajo estudio y para establecer las reglas básicas del intercambio de la misma entre entidades relacionadas con el mismo.

4.6 INTEGRACIÓN DE BASES DE DATOS BIOLÓGICAS

Según algunos estudios, el volumen de la información genómica se dobla cada 18 meses y, además, el número de bases de datos crece a una tasa anual de entorno al 5%. Todo ello tiene como resultado una dispersión enorme del conocimiento biológico que hace que el principal problema para el investigador sea la búsqueda de dicha información y para el bioinformático el diseño de un sistema capaz de proporcionarla.

Con el fin de dar respuesta a estos interrogantes se han desarrollado varias estrategias de integración que parten del hecho de que, aunque las bases de datos biológicas difieren en su estructura interna, implementación y ámbito biológico, la mayoría tienen una arquitectura de tres capas.

La estrategia de integración más sencilla es la basada en referencias. En este caso, para las entradas de una base de datos se proporciona una referencia a la información contenida sobre dicha entrada en otra base de datos diferente, de tal manera que, al presentar la información al usuario final esta referencia se convierte en un enlace a la otra base de datos. La principal ventaja de este enfoque es su sencillez de implementación aunque a costa de la actualización, fiabilidad y validez de la fuente externa. Por ejemplo, si buscamos en GenBank el registro correspondiente a la hemoglobina humana (ver Figura 4.3), veremos que aparecen una serie de enlaces y que al seguirlos nos llevan, en este caso, a la información recuperada de la base de datos taxonómica.

The screenshot illustrates the integration of biological data through references. On the left, the GenBank entry for the *Homo sapiens* hemoglobin (HBB) gene is shown. A red box highlights the taxonomic lineage: *Homo sapiens* (Human). On the right, the NCBI Taxonomy Browser interface is displayed, showing the same lineage and providing links to related information, such as the *Homo sapiens* (Human) entry and the *Homo sapiens* (Human) entry.

Figura 4.3. Integración basada en referencias

Un enfoque basado en servicios web soluciona estos inconvenientes a la vez que mantiene una simplicidad de implementación. En este caso, la integración con la base de datos externa se lleva a cabo por medio de la comunicación con un proceso remoto que expone los datos de la base de datos deseada. Manteniendo la interfaz de comunicación entre las bases de datos, cualquier cambio que se produzca en ambas será transparente para el otro extremo. Un ejemplo es la base de datos de dominios funcionales en proteínas PRODOM (<http://prodom.prabi.fr>), que ofrece una interfaz basada en servicios web para la consulta de información y la realización de operaciones básicas. La Figura 4.4 muestra el fichero de descripción de uno de sus servicios web:

```
<?xml version="1.0" encoding="UTF-8" ?>
<wsdl:definitions name="blastprodService" xmlns:soap="http://schemas.xmlsoap.org/wsdl/soap/" xmlns:tns="http://prodom.prabi.fr/blastprodom"
  xmlns:wsi="http://schemas.xmlsoap.org/wsdl/" xmlns:xsd="http://www.w3.org/2001/XMLSchema" targetNamespace="http://prodom.prabi.fr/blastprodom">
  <wsdl:documentation>Contact : Daniel Kahn, kahn@biomserv.univ-lyon1.fr Reference : Servant F, Bru C, Carrière S, Courcelle F, Gouzy J, Peyruc D, Kahn D (2002)
    ProDom: Automated clustering of homologous domains. Briefings in Bioinformatics, vol 3, no 3:246-251</wsdl:documentation>
  <wsdl:types>
    <xsd:schema elementFormDefault="qualified" targetNamespace="http://prodom.prabi.fr/blastprodom" xmlns:tns="http://prodom.prabi.fr/blastprodom"
      xmlns:xsd="http://www.w3.org/2001/XMLSchema">
      <xsd:element name="blastprod">
        <xsd:complexType>
          <xsd:sequence>
            <xsd:element name="program" type="xsd:string" />
            <xsd:element name="query" type="xsd:string" />
            <xsd:element name="param" type="xsd:string" />
          </xsd:sequence>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="blastprodResponse">
        <xsd:complexType>
          <xsd:sequence>
            <xsd:element name="blastprodOut" type="xsd:string" />
            <xsd:element name="blastprodImage" type="xsd:string" />
          </xsd:sequence>
        </xsd:complexType>
      </xsd:element>
    </xsd:schema>
  </wsdl:types>
  <wsdl:message name="blastprodRequestMsg">
    <wsdl:part name="inputWrapper" element="tns:blastprod" />
  </wsdl:message>
  <wsdl:message name="blastprodResponseMsg">
    <wsdl:part element="tns:blastprodResponse" name="outputWrapper" />
  </wsdl:message>
  <wsdl:portType name="blastprodomPortType">
    <wsdl:operation name="blastprod">

```

Figura 4.4. Integración a través de servicios web

Otra posibilidad es la llamada integración por vistas. Una vista, como su nombre indica, es una panorámica de la información que contiene una base de datos que está limitada por una serie de factores y que se consulta como si de información propia se tratara. Generalmente, las vistas se generan a partir de información de distintas tablas, de manera que el usuario que consulta la vista no necesita conocer la estructura de las tablas implicadas en la definición de la vista. En la Figura 4.5 se ve como la Vista A está formada a partir del subconjunto de las tablas de la base de datos, mientras que la Vista B permite el acceso a datos a otro subconjunto distinto.

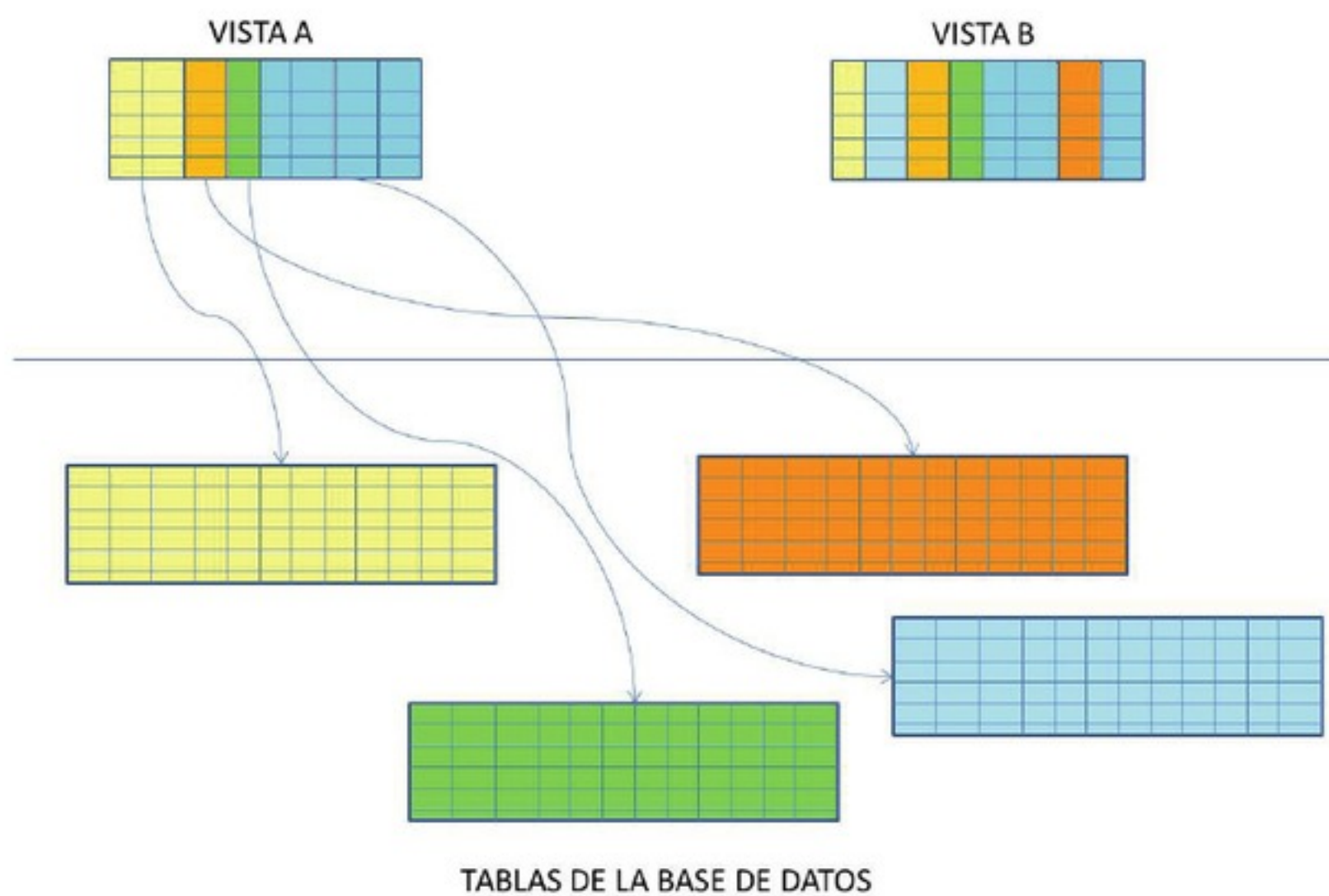


Figura 4.5. Integración por vistas

Finalmente, encontramos los almacenes de datos (*data warehouse*), cuyo objetivo es aglutinar, en una sola base de datos, un modelo de datos unificado que contenga la información de todas las fuentes de datos externas. El principal obstáculo que se encuentra es la naturaleza dinámica de la información y la dificultad de mantenimiento de la misma.

PRÁCTICA 1: DISEÑO DE BASES DE DATOS BIOLÓGICAS

En esta práctica sobre diseño de bases de datos biológicas propondremos un ejemplo simplificado de este tipo de bases de datos y abordaremos, con cierto nivel de detalle, el diseño de las mismas utilizando varias aproximaciones.

El ejemplo base será un sistema de información que almacenará datos sobre proteínas y las especies en que se encuentran, así como de las secuencias que las componen. Adicionalmente, será posible adjuntar, junto a las proteínas, anotaciones sobre las mismas.

5.1 DISEÑO RELACIONAL

La elaboración de un esquema Entidad-Relación es un proceso altamente creativo, por lo que resulta complejo establecer un procedimiento sistemático que garantice la resolución de cualquier tipo de problema. No obstante, la práctica demuestra que hay ciertas recomendaciones que simplifican el proceso de diseño.

En un diagrama Entidad-Relación encontramos, en la versión más sencilla, tres elementos:

- Entidades: son los objetos principales del modelo y recogen los conceptos del dominio bajo estudio. En este ejemplo, serían proteína, especie, secuencia y anotación.

- **Atributos:** representan características de las entidades del modelo y los hay de dos tipos: los identificadores (aquellos que permiten distinguir ocurrencias distintas de un ejemplar de la entidad) y los descriptores (el resto).
- **Relaciones:** son asociaciones entre una o más entidades, del mismo o de distinto tipo. Se caracterizan por la cardinalidad o número de entidades que se relacionan.

El diagrama Entidad-Relación¹ correspondiente al ejemplo que nos ocupa es el que se muestra en la Figura 5.1. En él, podemos encontrar seis entidades (PROTEIN, PROTEIN_SYNONYM, SPECIE, SEQUENCE, PROTEIN_INFO y PROTEIN_INFO_TYPE_CODE).

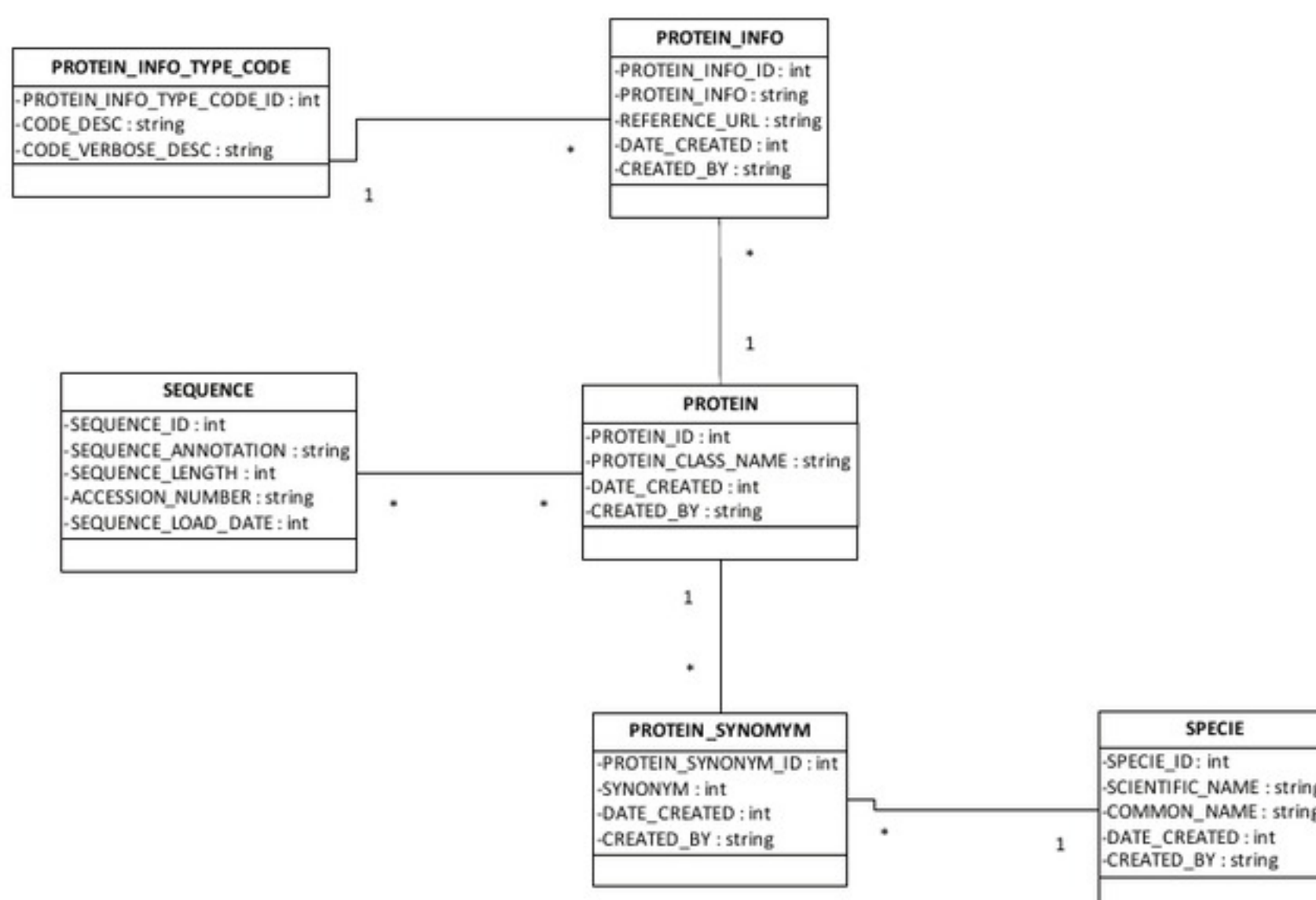


Figura 5.1. Diagrama Entidad-Relación

¹ La notación específica empleada para la construcción de este tipo de diagramas puede encontrarse en la Bibliografía.

Aunque no es el caso, podría ocurrir que aparecieran algunas restricciones semánticas que escapan a la capacidad de representación del modelo Entidad-Relación. Por este motivo, se extendió el modelo para dar lugar al modelo Entidad-Relación Extendido que, junto a lo anteriormente comentado incluye la generalización.

La generalización proporciona un mecanismo de abstracción que permite especializar una entidad, es decir, generar subtipos de un supertipo. Por ejemplo, una variación puede ser de tres tipos: inserción, delección o indel. En un diagrama Entidad-Relación estas relaciones quedarían representadas como:

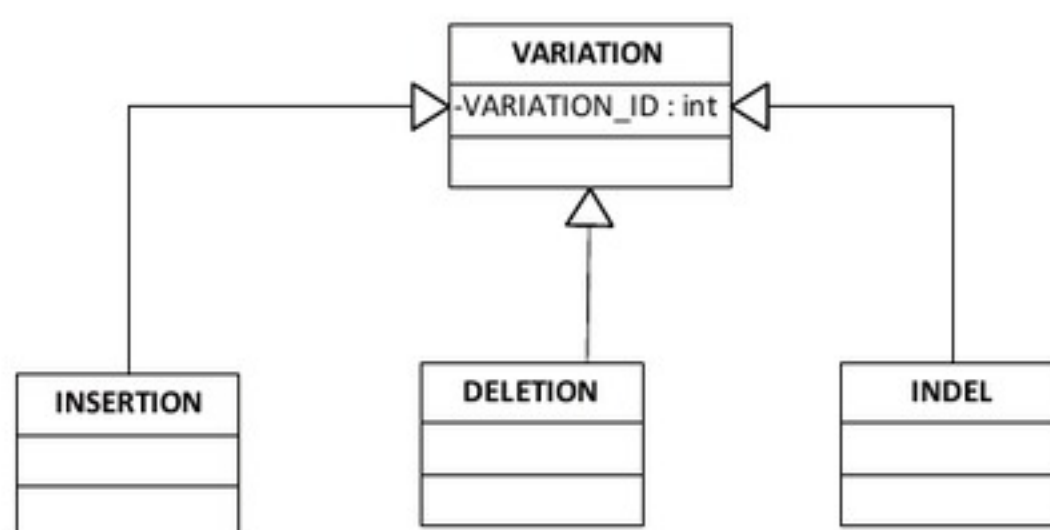


Figura 5.2. Ejemplo de generalización

Junto con esta relación de generalización aparecen las de totalidad/parcialidad y las de exclusividad/solapamiento. Una relación de generalización es total cuando no existe ninguna entidad del supertipo y no pertenece a alguno de los subtipos; en caso contrario, se dice que la generalización es parcial. Por otra parte, la generalización será exclusiva si una entidad pertenece únicamente a un subtipo. En el caso de las variaciones, se trata de una generalización total y exclusiva: todas las variaciones son una inserción, o una delección o un indel.

Una vez que se dispone del diagrama Entidad-Relación, el paso siguiente es transformar este modelo conceptual a un modelo relacional, es decir, obtener las tablas de la base de datos que contendrán la información que queremos almacenar. Las tres reglas básicas que debemos seguir son:

- Una entidad se transforma en una tabla cuya clave primaria es el identificador de la entidad.
- Las relaciones N:M se transforman en una tabla con una clave primaria formada por los identificadores de las entidades que relacionan.
- Las relaciones 1:N dan lugar o bien a una propagación de la clave o bien a una tabla.



A la hora de decidir si se propaga o no la clave, conviene hacer un esfuerzo en prever qué es lo que ocurrirá con la relación en el futuro. Si se cree posible que evolucione a una relación N:M, resulta recomendable transformar la relación 1:N en una tabla.

Si aplicamos estas reglas, obtenemos el esquema de tablas de la Figura 5.3:

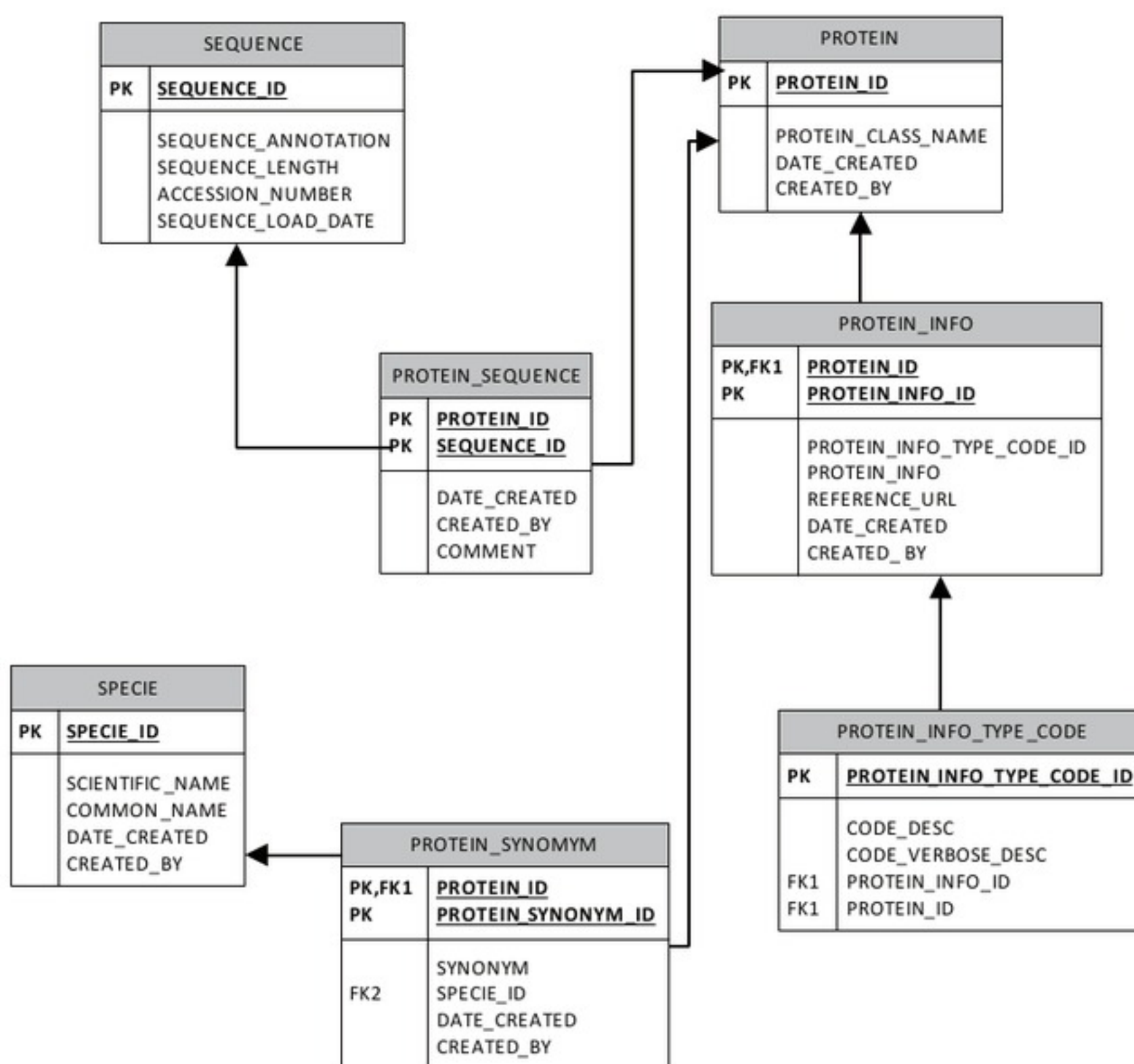


Figura 5.3. Ejemplo de esquema de tablas

5.2 DISEÑO XML

Un documento XML (*eXtended Markup Language*) es un documento de texto plano en el que pueden encontrarse etiquetas delimitadas por los signos de menor y mayor y que pueden anidarse entre sí. Algunas de estas etiquetas (o, más propiamente dicho, elementos) tienen atributos con valores especificados entre comillas.

La Figura 5.4 muestra un ejemplo de documento XML que sería válido para el modelo de datos de la base de datos biológica del ejemplo.

```
<proteins>
<protein created="05/09/2014" createdBy="darolmar">
  <specie value="Homo sapiens"/>
  <sequences>
    <sequence load_date="27/11/2003">
      <annotation>Anotación de ejemplo</annotation>
      <length>5</length>
      <accession>P.150308.1</accession>
    </sequence>
    <sequence load_date="01/12/2003">
      <annotation> notación de ejemplo 2</annotation>
      <length>8</length>
      <accession>P.150308.5</accession>
    </sequence>
  </sequences>
  <infos>
    <info created="13/04/1976" author="darolmar">
      <type>2</type>
      <value>Información sobre la proteína</value>
      <url>http://mi.host.es/protein?id=150508
    </info>
  </infos>
</protein>
</proteins>
```

Figura 5.4. Estructura de un documento XML



Si, durante el diseño XML, se duda entre incluir una determinada información como un elemento o como un atributo, generalmente se prefiere utilizar los atributos cuando la información es breve y sencilla, mientras que cuando la estructura de la información es más compleja, resulta recomendable recurrir a un elemento.

Un documento XML debe estar siempre bien formado, para lo que es requisito imprescindible que cumpla los siguientes requisitos:

- Estructura jerárquica de elementos: los documentos XML deben seguir una estructura estrictamente jerárquica respecto a las etiquetas que delimitan sus elementos:
 - Una etiqueta debe estar correctamente incluida en otra.
 - Los elementos con contenido deben estar correctamente cerrados. A diferencia de lo que ocurre en HTML, en XML a cada etiqueta se le asocia otra etiqueta de cierre.
- Etiquetas vacías: los elementos sin contenido se especifican de la forma `<elemento_sin_contenido/>`.
- Un solo elemento raíz: en un documento XML únicamente debe existir un elemento inicial.
- Delimitación de los valores de atributos: los valores de los atributos siempre van encerrados entre comillas.
- Tipo de letras: XML es sensible a las mayúsculas y las minúsculas.

Además, para una determinada aplicación, el documento será válido si es posible comprobar que cumple con las reglas semánticas recogidas en otro documento: un DTD o un XMLSchema.

Un DTD (*Document Type Definition*) es un documento con una sintaxis especial que se emplea para validar e interpretar el contenido de un documento XML. En la Figura 5.5 podemos ver un posible DTD para el ejemplo que nos ocupa:

```
<!DOCTYPE proteins [  
  <!ELEMENT (proteins)*>  
  <!ELEMENT protein (specie, sequences, infos)>  
  <!ATTLIST protein created CDATA>  
  <!ATTLIST protein createdBy CDATA>  
  <!ELEMENT specie EMPTY>  
  <!ATTLIST specie load_data CDATA>  
  <!ELEMENT sequences (sequence)*>  
  <!ELEMENT sequence (annotation, length, accession)>  
  <!ATTLIST sequence load_date CDATA>  
  <!ELEMENT annotation (#PCDATA)>  
  <!ELEMENT length (#PCDATA)>  
  <!ELEMENT accession (#PCDATA)>  
  <!ELEMENT infos (info)*>  
  <!ELEMENT info (type, value, url)>  
  <!ATTLIST info created CDATA>
```



```

<!ATTLIST info author CDATA>
<!ELEMENT type (#PCDATA)>
<!ELEMENT value (#PCDATA)>
<!ELEMENT url (#PCDATA)>
]>

```

Figura 5.5. DTD de ejemplo

En la definición del elemento, los valores entre paréntesis significan lo siguiente (ver Tabla 5.1):

Protein	Exactamente un elemento protein
Protein?	Ninguno o un elemento protein
Protein+	Uno o más elementos protein
Protein*	Ninguno, uno o más elementos protein

Tabla 5.1. Indicadores de cardinalidad

Además, entre paréntesis se indica el contenido posible de un elemento:

specie, sequences	Tiene un elemento specie y a continuación un elemento sequences
sequences infos	Tiene un elemento sequences o un elemento infos

Tabla 5.2. Contenido de un elemento

Y, finalmente, se especifican los atributos para cada uno de los elementos, para lo que se utiliza la expresión:

```
<!ATTLIST elemento atributo tipo modo>
```

donde:

Elemento	Es el nombre del elemento al que pertenece el atributo
Atributo	Es el nombre del atributo
Tipo	Es el tipo del atributo, que puede ser: CDATA: cadena de caracteres (val1 ... valN): tipo enumerado
Modo	Modos de actuar: #REQUIRED: obligatorio #IMPLIED: opcional "valor": valor por defecto (opcional) #FIXED "valor": si el atributo aparece, tendrá obligatoriamente el valor "valor"

Tabla 5.3. Definición de un atributo



XML.

El DTD es opcional y solo se debe utilizar si la aplicación exige la validación del documento. Si únicamente es necesario que esté formado, puede ser recomendable no utilizar DTD, por ejemplo, porque se esté trabajando con un conjunto pequeño de documentos

Como hemos visto, un DTD tiene una sintaxis muy farragosa, sobre todo cuando la estructura del documento XML se complica. Además, tiene el inconveniente de que es necesario aprender un idioma distinto del XML.

Para resolver estos inconvenientes se pensó en definir la estructura de la información de un documento XML y sus reglas de validación en otro documento XML, el XMLSchema.

La Figura 5.6 muestra el XMLSchema equivalente al DTD anterior. En la Bibliografía se encontrará un estudio detallado de cómo definir documentos XMLSchema, aquí únicamente nos interesa llamar la atención sobre la sencillez de comprensión de esta alternativa, si se la compara con el DTD.

```
<?xml version="1.0"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
targetNamespace="http://www.mibiosoftware.es"
  xmlns="http://www.mibiosoftware.es"
  elementFormDefault="qualified">
<xs:element name="proteins" type="TipoListaProteinas"/>
<xs:complexType name="TipoListaProteinas">
  <xs:sequence>
    <xs:element name="protein" type="TipoProteina"
maxOccurs="unbounded"/>
  </xs:sequence>
</xs:complexType>
<xs:complexType name="TipoProteina">
  <xs:sequence>
    <xs:element name="specie" type="TipoSpecie" maxOccurs="1"/>
    <xs:element name="sequences" type="TipoListaSecuencias"
maxOccurs="1"/>
    <xs:element name="infos" type="TipoListaInfo"
maxOccurs="1"/>
  </xs:sequence>
  <xs:attribute name="created" type="xs:date"/>
</xs:complexType>
</xs:schema>
```



```
<xs:attribute name="createdBy" type="xs:string"/>
</xs:complexType>
<xs:complexType name="xs:string">
  <xs:attribute name="value" type="xs:string"/>
</xs:complexType>
<xs:complexType name="TipoListaSecuencias">
  <xs:sequence>
    <xs:element name="sequence" type="TipoSecuencia"
maxOccurs="unbounded"/>
  </xs:sequence>
</xs:complexType>
<xs:complexType name="TipoSecuencia">
  <xs:sequence>
    <xs:element name="annotation" type="xs:string" maxOccurs="1"/>
    <xs:element name="length" type="xs:positiveInteger"
maxOccurs="1"/>
    <xs:element name="accession" type="xs:string" maxOccurs="1"/>
  </xs:sequence>
  <xs:attribute name="load_date" type="xs:date"/>
</xs:complexType>
<xs:complexType name="TipoListaInfo">
  <xs:sequence>
    <xs:element name="info" type="TipoInfo"
maxOccurs="unbounded"/>
  </xs:sequence>
</xs:complexType>
<xs:complexType name="TipoInfo">
  <xs:sequence>
    <xs:element name="type" type="xs:postiveInteger"
maxOccurs="1"/>
    <xs:element name="value" type="xs:string" maxOccurs="1"/>
    <xs:element name="url" type="xs:string" maxOccurs="1"/>
  </xs:sequence>
  <xs:attribute name="created" type="xs:date"/>
  <xs:attribute name="author" type="xs:date"/>
</xs:complexType>

</xs:schema>
```

Figura 5.6. Ejemplo anterior con XMLSchema



6

PRINCIPALES BASES DE DATOS GENÓMICAS

Como ya se dijo al principio de este capítulo, existe una gran variedad de bases de datos genómicas. Queda fuera del ámbito de este libro ofrecer una descripción detallada de todas y cada una de las mismas, por lo que nos centraremos únicamente en aquellas que vayan a emplearse en los casos prácticos propuestos a lo largo de esta obra que, por otra parte, son las de uso más común.

6.1 GENBANK

GenBank es una colección pública de secuencias de nucleótidos anotadas que incluye secuencias de ARNm con regiones codificantes, ADN genómico correspondientes a uno o varios genes y ARN ribosómico.

La información contenida en GenBank se organiza en dos divisiones o categorías cuyo objetivo es simplificar la búsqueda: en la primera división (*orgasimal*) se incluyen bases de datos de secuencias derivadas de organismos específicos; por su parte, bajo la categoría funcional (*functional*) se engloban bases de datos de secuencias de distintos tipos. Cada registro de GenBank pertenece a una única división. La Tabla 6.1 y la Tabla 6.2 muestran la situación actual de las bases de datos de secuencias, clasificadas por divisiones:

Base de datos	División
BCT	Secuencias de bacterias
PRI	Secuencias de primates
ROD	Secuencias de roedores
MAM	Secuencias de otros mamíferos
VRT	Secuencias de otros vertebrados
INV	Secuencias de invertebrados
PLN	Secuencias de plantas y hongos
VRL	Secuencias de virus
PHG	Secuencias de fagos
RNA	Secuencias de ARN estructural
SYN	Secuencias sintéticas y quiméricas
UNA	Secuencias sin anotar

Tabla 6.1. Bases de datos de la división de organismos

Base de datos	División
EST	Etiquetas EST
STS	Sitios STS
GSS	Secuencias de encuestas de genomas
HTG	Secuencias de alto rendimiento

Tabla 6.2. Bases de datos de la división funcional

Esta base de datos está orientada a almacenar datos primarios de secuencias. Cuando se envía una nueva secuencia para su incorporación a GenBank, el NCBI lleva a cabo un control de calidad básico y lo notifica al remitente pero en ningún caso se revisa y se contrasta la información: el responsable de la misma es el propio remitente. De hecho, a pesar de que se insiste a los autores para que actualicen sus registros con nuevos datos sobre las secuencias o sus anotaciones, lo cierto es que en la práctica es habitual encontrar registros sin actualizar.

Al ser GenBank un archivo genómico e incluir todos los datos de secuencias que se envían, es frecuente encontrar varias entradas distintas para un mismo loci. Las diferencias en los envíos reflejan variaciones genéticas entre individuos u organismos y analizar estas diferencias es una manera de identificar polimorfismos de un solo nucleótido.

GenBank intercambia diariamente información con dos *partners* del INSDC (International Nucleotide Sequence Database Collaboration): el EBI (European Bioinformatics Institute) de EMBL (European Molecular Biology Laboratory) y el banco de ADN del DDBJ (Data Bank of Japan). La mayoría de datos de secuencias depositados en las bases de datos del INSDC procede de laboratorios de generación de secuencias, en parte porque muchas publicaciones exigen el depósito previo con el fin de que el *accession* se incluya en el artículo.

Si parte de una secuencia de nucleótidos de GenBank codifica una proteína, se anota una traducción conceptual llamada secuencia o región de codificación (CDS) y se le asigna un identificador o *accession* que caracterizará al registro que se añade y se encuentra enlazado con un registro en la base de datos de proteínas del NCBI o en la sección TrEMBL de la base de datos Uniprot.

6.1.1 Formato del registro

El formato de los registros de GenBank es uno de los más comúnmente utilizados para describir información biológica, por lo que lo estudiaremos con cierto nivel de detalle.

Un registro en GenBank contiene un fichero plano (GBFF, GenBank *Flat File*) cuyo contenido se divide en tres partes: cabecera, que contiene descriptores que se aplican a todo el registro; características o anotaciones sobre el registro; y, la secuencia propiamente dicha. Al final de cada registro, encontramos dos barras ‘//’ que indican el final del fichero.

Para explicar cómo se estructura la información en un registro de GenBank, utilizaremos el siguiente ejemplo (ver Figura 6.1):

```
LOCUS      V00505          1976 bp  DNA   linear  PRI 14-NOV-2006
DEFINITION Human gene for delta-globin.
ACCESSION  V00505
VERSION    V00505.1 GI:30510
KEYWORDS   delta globin; germ line; globin.
SOURCE     Homo sapiens (human)
ORGANISM   Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates;
            Haplorrhini;
            Catarrhini; Hominidae; Homo.
```

```

REFERENCE 1 (bases 1 to 1976)
AUTHORS Spritz,R.A., DeRiel,J.K., Forget,B.G. and Weissman,S.M.
TITLE Complete nucleotide sequence of the human delta-globin
gene
JOURNAL Cell 21 (3), 639-646 (1980)
PUBMED 7438204
COMMENT KST HSA.DELGLOBIN.
FEATURES             Location/Qualifiers
     source             1..1976
                        /organism="Homo sapiens"
                        /mol_type="genomic DNA"
                        /db_xref="taxon:9606"
     prim_transcript 123..1763
     exon               123..265
                        /number=1
     CDS                join(173..265,394..615,1505..1633)
                        /codon_start=1
                        /product="delta globin"
                        /protein_id="CAA23763.1"
                        /db_xref="GI:30511"
                        /db_xref="GDB:119298"
                        /db_xref="GOA:P02042"
                        /db_xref="HGNC:4829"
                        /db_xref="InterPro:IPR000971"
                        /db_xref="InterPro:IPR002337"
                        /db_xref="InterPro:IPR009050"
                        /db_xref="InterPro:IPR012292"
                        /db_xref="PDB:1SHR"
                        /db_xref="PDB:1SI4"
                        /db_xref="UniProtKB/Swiss-Prot:P02042"
                        /translation="MVHLTPEEKTAVNALWGKVNVDVAVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLNLKGTFSQLSELHCDKLHVDPE
NFRLLGNVLVLCVLAARNGKEFTPMQAAVQKVVAGVANALAHKYH"
     intron             266..393
                        /number=1
     exon               394..615
                        /number=2
     intron             616..1504
                        /number=3
     exon               1505..1763
                        /number=3
ORIGIN
      1 aatgaagggt catttttcat tctcacaac taatgaaacc ctgcttatct
taaaccaacc

```



```
61 tgctcactgg agcagggagg acaggaccag cataaaaggc agggcagagt
cgactgttgc
121 ttacactttc ttctgacata acagtgttca ctagcaacct caaacagaca
ccatgggtgca
181 tctgactcct gaggagaaga ctgctgtcaa tgccctgtgg ggcaaagtga
acgtggatgc
241 agttggtggt gaggccctgg gcaggttgg atcaaggtta taagagaggg
tcaaggaggg
301 aaatggaaac tgggcatgtg tagacagaga agactcttgg gtttctgata
ggcactgact
361 ctctgtccct tgggctgttt tectaccctc agattactgg tggctaccc
ttggacccag
421 aggttctttg agtcctttgg ggatctgtcc tctcctgatg ctgttatggg
caaccctaag
481 gtgaaggctc atggcaagaa ggtgctaggt gccttttagtg atggcctggc
tcacctggac
541 aacctcaagg gcactttttc tcagctgagt gagctgcact gtgacaagct
gcacgtggat
601 cctgagaact tcagggtgag tccaggagat gcttcacttt tctcttttta
ctttctaate
661 ttacattttg gttctttttac ctacctgctc ttctcccaca tttttgtcat
tttactatat
721 tttatcattt aatgcttcta aaattttggt atttttttat ttaaaaattc
tgcatttttt
781 ccttctcac aatcttgcta ctctaaatta tttaatatcc tgtctttctc
tcccacccc
841 ctcccttcat ttttcttct ctaacaacaa ctcaaattat gcataccagc
tctcacctgc
901 taattttgca cttagaataa tccttttgtc tctccacatg ggtatgggag
aggctccaac
961 tcaaagatga gaggcataga atactgtttt agaggctata aatcatttta
caataaggaa
1021 taattggaat ttataaatt ctgtagtaaa tggaatggaa aggaaagtga
atatttgatt
1081 atgaaagact aggcagttac actggaggtg gggcagaagt cgttgctagg
agacagcca
1141 tcatcacact gatttatcaa ttcaatttgt atctattaat ctgtttatag
taattaattt
1201 gtatatgcta tatacacata caaaattaaa actaatttgg aattaatttg
tatatagtat
1261 tatacagcat atatgtacat atatagacta catgctagtt aagtacatag
aggatgtgtg
1321 tgtatagata tatgttatat gtatgcattc atatatgtac ttatttatgc
tgatgggaat
1381 aacctgggga tcagttttgt ctaagatttg ggcagaaaaa aatgggtgtt
```

```

ggctcagttc
  1441 tcagaagcca gtctttatct ctctgttaac catatgcatg tatctgccta
cctcttctcc
  1501 gcagctcttg ggcaatgtgc tgggtgtgtg gctggcccg c aactttggca
aggaattcac
  1561 cccacaaatg caggctgcct atcagaaggt ggtggctggt gtggctaata
ccttggctca
  1621 caagtaccat tgagatcctg gactgtttcc tgataaccat aagaagaccc
tatttccta
  1681 gattctatct tctgaacttg ggaacacaat gcctacttca agggatatggc
ttctgcctaa
  1741 taaagaatgt tcagctcaac ttcttgatta atttcaacta tttcattttt
ttgtccaggt
  1801 gtgtaagaag gttcctgagg ctctacagat agggagcact tctttatctt
acaaagagta
  1861 catgggaaaa gagaaaagca agggaaccgt acaaggcatt aatgggtgac
acttctacct
  1921 ccaaagagca gaaattatca agaactcttg atacaaagat aatactggca
ctgcag
//

```

Figura 6.1. Ejemplo de registro GenBank

6.1.2 Cabecera

Esta parte del registro es específica de la base de datos. Las diferentes bases de datos existentes en GenBank no están obligadas a incluir la misma información en este segmento del registro, sino que existen algunas pequeñas variaciones a pesar de que se procura que todas ellas contengan la misma información.

La primera línea de la cabecera es el LOCUS:

LOCUS	V00505	1976 bp	DNA	linear	PRI 14-NOV-2006
-------	--------	---------	-----	--------	-----------------

En primer lugar, encontramos el nombre del locus, que es único en toda la base de datos y que identifica al registro dentro de la misma. El nombre del locus recibe el nombre de identificador de acceso o accessionid y se trata de un campo alfanumérico con todas las letras mayúsculas. En este caso, se trata de la secuencia V00505.



En muchas de las herramientas que veremos a lo largo del libro, podremos trabajar directamente con la secuencia o bien referenciarla por su *accession*.

El segundo elemento es la longitud de la secuencia, que está comprendida entre 1 y 350000 bp (pares de bases) por registro. La secuencia V00505 está compuesta por 1976 bases.

A continuación, encontramos el tipo de molécula que podrá ser DNA, RNA, tRNA, rRNA, mRNA o uRNA.

El cuarto elemento de la línea de LOCUS indica la estructura de la molécula (*linear* o *circular*) mientras que el quinto es un campo de tres letras que establece el tipo de división (ver Tabla 1.2 y Tabla 1.3). Finalmente, se encuentra la fecha en que publicó la última actualización de la secuencia.

Resumiendo, según el contenido de la línea LOCUS podemos afirmar que la secuencia V00505 pertenece a un primate, que se trata de una secuencia de ADN de estructura lineal actualizada por última vez el 14 de noviembre de 2006.

La siguiente línea de la cabecera es la de definición, cuyo propósito es resumir la información biológica que contiene el registro.

```
DEFINITION Human gene for delta-globin.
```

El *accession*, en la tercera línea de la cabecera, constituye la clave primaria por la que se referencia al registro en la base de datos. Todos los registros tienen una única línea *accession*:

```
ACCESSION V00505
```

No obstante, es cierto que en algunos es posible encontrar más de un *accession*. Si este es el caso, habrá un *accession* primario y el resto serán secundarios. El significado de estos últimos ha ido cambiando a lo largo de los años.

Tras el *accession*, aparece la línea de versión que, como su nombre indica, ofrece información sobre la versión de la secuencia junto con un identificador del gen (gi, *geninfo identifier*). Estos identificadores se asocian a una secuencia de nucleótidos única. Si cambia la secuencia, se incrementa el número de versión en una unidad y el gi también se modificará.

Es posible añadir palabras clave al registro. Sin embargo, puesto que tradicionalmente ha sido un campo de texto libre sobre el que no se ha llevado ningún tipo de control, el NCBI desaconseja su utilización, a pesar de que la incluye si aparece en la petición de creación de un nuevo registro.

```
KEYWORDS  delta globin; germ line; globin.
```

La línea de fuente contiene el nombre común o científico del organismo al que pertenece la secuencia. En la actualidad, el NCBI está haciendo un esfuerzo para obtener esta información taxonómica directamente desde las bases de datos correspondientes.

```
SOURCE  Homo sapiens (human)
ORGANISM Homo sapiens
          Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
Euteleostomi;
          Mammalia; Eutheria; Euarchontoglires; Primates;
Haplorrhini;
          Catarrhini; Hominidae; Homo.
```

Por otra parte, cada registro de GenBank puede tener una o más referencias o citas:

```
REFERENCE  1 (bases 1 to 1976)
AUTHORS  Spritz,R.A., DeRiel,J.K., Forget,B.G. and Weissman,S.M.
TITLE    Complete nucleotide sequence of the human delta-globin
gene
JOURNAL  Cell 21 (3), 639-646 (1980)
PUBMED   7438204
```

Finalmente, la última línea de la cabecera es la línea de comentarios en la que se incluye información variada (descriptores).

```
COMMENT   KST HSA.DELGLOBIN.
```


6.1.3 Sección de características

La tabla de características es la representación de la información biológica más importante contenida en el registro de GenBank y hace referencia a anotaciones sobre una parte de la secuencia, mientras que las anotaciones sobre la secuencia completa reciben el nombre de descriptores.



Se puede obtener una lista completa de las características incluidas en un registro de GenBank, así como de la información asociada a las mismas en <ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>.

La característica SOURCE es la única que está presente en todos los registros de GenBank. Todas las características tienen una serie de calificadores permitidos, algunos de los cuales son obligatorios (como */organism* para SOURCE).

```
source      1..1976
            /organism="Homo sapiens"
            /mol_type="genomic DNA"
            /db_xref="taxon:9606"
prim_transcript 123..1763
exon        123..265
            /number=1
```

Otra de las características significativas es la característica CDS, que contiene las instrucciones de obtención de una secuencia siguiendo las coordenadas que se indican en el rango. Es posible que haya que unir varias secuencias, en cuyo caso, se utiliza la palabra clave *join*.

```
CDS         join(173..265,394..615,1505..1633)
            /codon_start=1
            /product="delta globin"
            /protein_id="CAA23763.1"
            /db_xref="GI:30511"
            /db_xref="GDB:119298"
            /db_xref="GOA:P02042"
            /db_xref="HGNC:4829"
            /db_xref="InterPro:IPR000971"
            /db_xref="InterPro:IPR002337"
            /db_xref="InterPro:IPR009050"
```

```

/db_xref="InterPro:IPR012292"
/db_xref="PDB:1SHR"
/db_xref="PDB:1SI4"
/db_xref="UniProtKB/Swiss-Prot:P02042"
/translation="MVHLTPEEKTAVERNALWGKVNVDAGVGGEALGRLL
VVYPWTQRFFESFGDLSSPDAMGNPKVKAHGKKVLGAFSDGLAHLD
NLKGTFSQLSEHLCDEKLVHDPENFRLLGNVLVCVLAARNFGKEFTPM
QAAYQKVAVGVALAHKYH"

```

Este ejemplo también muestra el uso de referencias cruzadas con otras bases de datos (calificador `/db_xref`). El contenido de este calificador se divide en dos partes: la primera se refiere a la base de datos con la que se cruza la referencia mientras que la segunda, que está separada de la primera por dos puntos, es el identificador de la secuencia en dicha base de datos.

6.1.4 Sección ORIGIN

La sección ORIGIN contiene la secuencia de bases en filas de 60 bases agrupadas en columnas de 10 nucleótidos y precedidas de un número que indica la posición, dentro del total de nucleótidos, que ocupa la primera base de la línea correspondiente. De esta manera, resulta más sencillo localizar un fragmento de la secuencia para, por ejemplo, averiguar la secuencia de una CDS.



El contenido de esta sección incluye números y letras y no puede utilizarse directamente en las herramientas de análisis de secuencias. En su lugar, hay que obtener la secuencia en cuestión en formato FASTA, lo que puede conseguirse fácilmente seleccionando esta opción del menú desplegable.

6.2 REFSEQ

La base de datos RefSeq (*Reference Sequence*) es una base de datos secundaria y revisada de secuencias de ADN, ARN y proteínas construida por el NCBI. A diferencia de GenBank, RefSeq contiene un único registro por cada molécula biológica para los organismos más importantes, desde los virus hasta bacterias y eucariotas, de los que se tiene suficiente información.

Para cada modelo de organismo, RefSeq intenta proporcionar una serie de registros separados y enlazados entre sí. El formato de un registro de RefSeq es similar al del GenBank excepto porque en RefSeq el *accession* incluye un subrayado y en el campo COMMENT se especifica el estado del registro de RefSeq (ver Tabla 6.3).

Código	Descripción
MODEL	El registro ha sido proporcionado por el <i>pipeline</i> de anotación de genomas del NCBI y no es una revisión entre ejecuciones
INFERRED	El registro es una predicción obtenida del análisis del genoma, pero no hay ninguna evidencia empírica
PREDICTED	El registro todavía no se ha revisado, aunque algunos aspectos se han predicho
PROVISIONAL	El registro todavía no se ha revisado
REVIEWED	El registro ha sido revisado por el personal del NCBI o algún colaborador. Algunos registros pueden incluir anotaciones
VALIDATED	El registro ha pasado con éxito la revisión inicial pero está pendiente de una revisión final en la que es posible incluir información funcional
WGS	El registro representa un genoma completo y está pendiente de una revisión individual o de revisiones entre actualizaciones del genoma.

Tabla 6.3. Códigos de estado de un registro de RefSeq

La Tabla 6.4 resume las principales diferencias entre RefSeq y GenBank:

GenBank	RefSeq
No es curada	Curada
El autor envía las secuencias	Es el NCBI quien genera los registros a partir de datos existentes
Solamente el autor puede revisar las secuencias que envía	El NCBI revisa la información a medida que se va generando
Es posible encontrar varios registros para el mismo loci	Un único registro para cada molécula de los organismos superiores
Probabilidad de encontrar registros con información contradictoria	
No hay límite en las especies incluidas	Limitado a organismos modelo
Datos intercambiados con los miembros del INSDC	Información obtenida únicamente de las bases de datos del NCBI
Enlaces a las proteínas identificadas	Enlaces a las proteínas y a los transcritos identificados

Tabla 6.4. Comparación entre RefSeq y GenBank

6.3 UNIPROT

UniProt (*Universal Protein Resource*) es una base de datos de secuencias de proteínas que se formó como resultado de la unificación de otras tres bases de datos: Swiss-prot (proteínas mejor anotadas por expertos), TrEMBL (proteínas que no están en Swiss-Prot encontradas automáticamente) y PIR-PSD (proteínas anotadas por expertos).

6.4 PDB

Protein Data Bank (PDB, <http://www.rcsb.org/pdb/home/home.do>) es un repositorio de información sobre proteínas y otras macromoléculas biológicas importantes. Contiene información sobre su estructura 3D obtenida por cristalografía de rayos X y RMN.

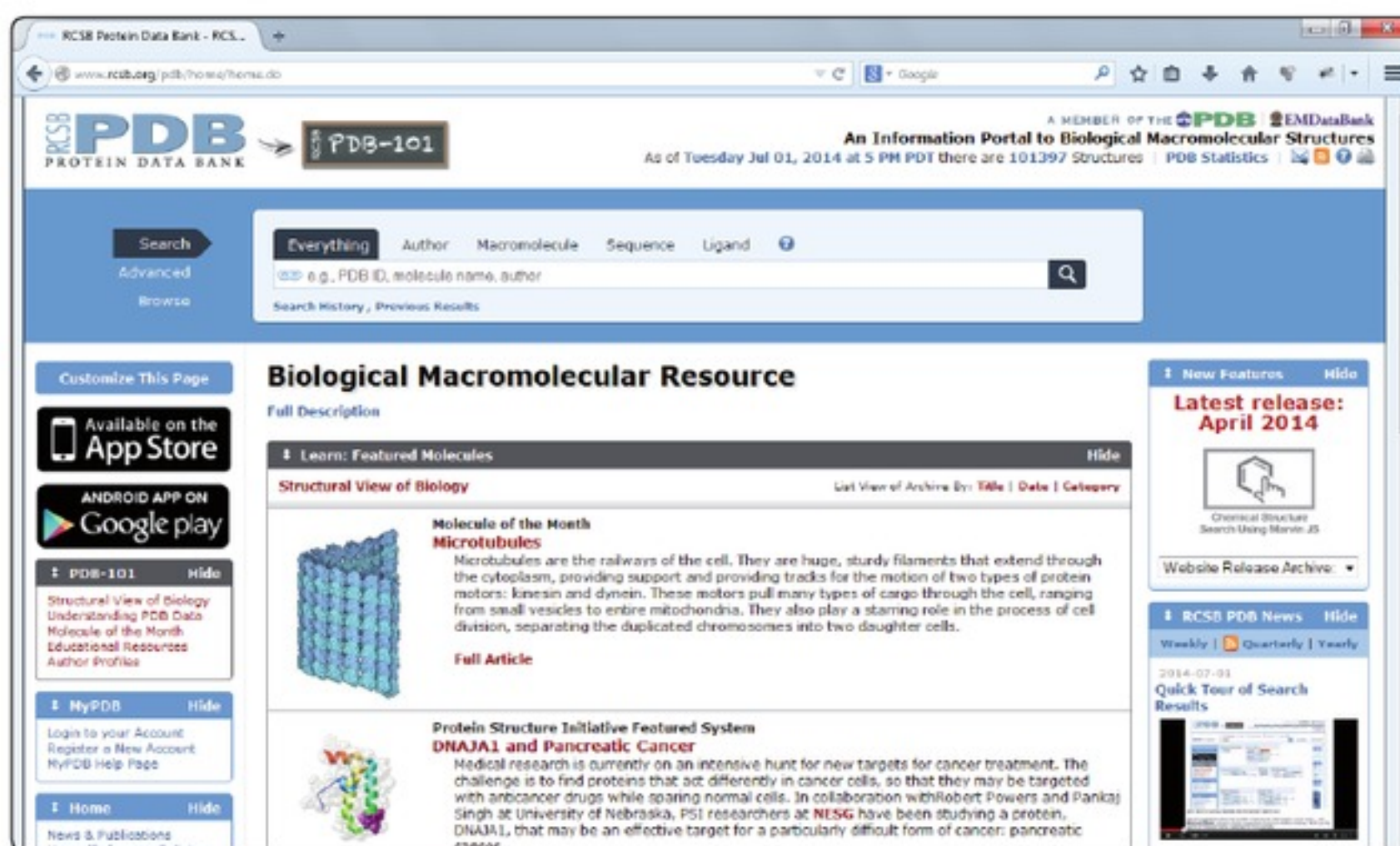


Figura 6.2. Página principal de PDB

A través del portal web de PDB, podemos encontrar la siguiente información:

- **Ficheros en formato PDB:** básicamente, un fichero PDB es una serie de coordenadas de moléculas biológicas que, en conjunto, representan la estructura de las mismas. Este formato de fichero se describirá con mayor nivel de detalle en epígrafes posteriores.
- **Visualización de estructuras:** además de los ficheros en formato PDB, también es posible encontrar herramientas que permitan la navegación y la visualización de la información contenida en dichos ficheros.



Cualquier herramienta software que lea ficheros PDB debe ser capaz de reconstruir los enlaces químicos basándose en una serie de reglas (por ejemplo, que un enlace está formado por dos puntos separados en el espacio real por 1,5 Å). Sin embargo, se trata de convenciones de cumplimiento no obligatorio, por lo que el programador tiene libertad para elegir y, por tanto, la misma molécula puede presentar un aspecto diferente en herramientas distintas.

6.4.1 Formato del registro

Todos los ficheros PDB son ficheros de texto, de extensión variable, cuyas líneas constan de 80 caracteres, de los cuales los 6 primeros corresponden al nombre del registro. Cada línea puede contener un tipo de registro diferente y cada tipo de registro se caracteriza por un descriptor de la información contenida en el mismo y se divide en los campos *Overview*, *Record Format*, *Details*, *Verification/Validation/Value Authority Control*, *Relationship to other record types*, *Examples* y *Known Problems*. Las columnas que no se utilicen, se dejan en blanco.

6.4.2 Tipos de registros

Los registros se agrupan en categorías basándose en la frecuencia de aparición del tipo de registro en una entrada PDB dada.

La primera de las categorías es la de aparición única en una sola línea (OTSL, *One Time, Single Line*). Las líneas correspondientes a estos registros solamente aparecen una vez en todo el fichero y son las de la Tabla 6.5.

Registro	Descripción
CRYST1	Parámetros de celda-unidad, grupo especial y Z
END	Último registro del fichero
HEADER	Primera línea del fichero que contiene el código del identificador PDB ID, la clasificación y la fecha de depósito
NUMMDL	Número de modelos
MASTER	Registro de control para la búsqueda automática y funciones de biblioteca
ORIGXn	Transformación de coordenadas ortogonales a las coordenadas actuales (n = 1, 2, or 3)
SCALEn	Transformación de coordenadas ortogonales a las coordenadas cristalográficas (n = 1, 2, or 3)

Tabla 6.5. Tipos de registro OTSL

Otra de las categorías es la de los registros que aparecen una vez en todo el fichero pero que, a diferencia de los registros OTSL, ocupan varias líneas del fichero. Son los registros OTML (*One Time, Multiple Lines*).

Tipo de registro	Descripción
AUTHOR	Lista de personas que han contribuido a su edición
CAVEAT	Indicador de error severo
COMPND	Descripción del contenido macromolecular del fichero PDB
EXPDTA	Técnicas experimentales empleadas para la determinación de la estructura
MDLTYP	Anotaciones adicionales correspondientes a las coordenadas indicadas en el registro
KEYWDS	Lista de palabras clave que describen la macromolécula
OBSLTE	El registro ha sido eliminado porque se ha reemplazado otro cuyo ID se indica aquí
SOURCE	Fuente biológica de la macromolécula
SPLIT	Lista de las entradas PDB que componen una molécula compleja
SPRSDE	Lista de identificadores de versiones anteriores y la última (aparece en primer lugar) con su fecha
TITLE	Descripción del experimento representado en esta entrada

Tabla 6.6. Registros de tipo OTML

La tercera de las categorías son los registros MTOL (*Multiple Times, One Line*). Se trata de registros que aparecen varias veces en el fichero, a menudo en grupos donde la información no está relacionada desde el punto de vista lógico pero que, por alguna razón, se presenta junta.

Tipo de registro	Descripción
ANISOU	Factores de temperatura anisotrópica
ATOM	Coordenadas atómicas de grupos estándar
CISPEP	Identificación de residuos de péptidos en conformación CIS
CONNECT	Registros de conectividad
DBREF	Referencia de la entrada en la base de datos de secuencias
HELIX	Identificación de la subestructura helicoidal
HET	Identificación de grupos o restos no estándar
HETATM	Coordenadas atómicas de grupos heterogéneos
LINK	Identificación de enlaces entre residuos
MODRES	Identificación de modificaciones en residuos estándar
MATRIXn	Transformaciones que expresan simetría no cristalográfica (n = 1, 2, or 3)
REVDAT	Fecha de la revisión e información relacionada
SEQADV	Identificación de conflictos entre PDB y la base de datos de secuencias
SHEET	Identificación de subestructura lámina b
SSBOND	Identificación de puentes disulfuro

Tabla 6.7. Registros de tipo MTSL

Por otra parte, encontramos los registros con múltiples ocurrencias y que ocupan varias líneas del fichero (MTML, *Multiple Time, Multiple Line*).

Tipo de registro	Descripción
FORMUL	Fórmula química de grupos no estándar
HETNAM	Coordenadas atómicas de grupos heterogéneos
HETSYN	Sinónimos de grupos heterogéneos
SEQRES	Secuencia primaria de los residuos que forman la macromolécula
SITE	Identificación de grupos que forman sitios importantes

Tabla 6.8. Registros de tipo MTML

Así mismo, es posible encontrar también grupos de registros (ver Tabla 6.9). Los registros ENDMDL y MODEL agrupan registros de tipo ATOM, HETATM, ANISOU y TER.

Tipo de registro	Descripción
ENDMDL	Registro de fin de modelo para estructuras múltiples en una única coordenada de registro
MODEL	Especificación del número de modelo para estructuras múltiples en una única coordenada de registro
TER	Terminador de cadena

Tabla 6.9. Registros de tipo agrupación

Finalmente, encontramos otros tipos de registros que no pueden clasificarse en ninguno de los grupos anteriores y que presentan una estructura muy particular (ver Tabla 6.10).

Tipo de registro	Descripción
JRNL	Cita que define el conjunto de coordenadas
REMARK	Apuntes generales, estructurados en formato libre

Tabla 6.10. Otros tipos de registros

6.4.3 Estructura del fichero

En un fichero PDB es muy importante el orden en que aparecen los registros y la sección en que aparecen. La Tabla 6.11 resume esta información:

Sección	Descripción	Tipos de registro
Title	Resumen y apuntes descriptivos	HEADER, OBLTE, TITLE, SPLIT, CAVEAT, COMPND, SOURCE, KEYWDS, EXPDTA, NUMMDL, MDLTYP, AUTHOR, REVDAT, SPRSDE, JRNL

Remark	Comentarios sobre las anotaciones de la entrada que se ha considerado importante indicar con mayor nivel de detalle que en un registro estándar	REMARKs 0-999
Primary structure	Secuencia de péptidos o nucleótidos y su relación con la secuencia PDB y la encontrada en la base de datos de secuencias	DBREF, SEQADV, SEQRES MODRES
Heterogen	Descripción de grupos no estándar	HET, HETNAM, HETSYN, FORMUL
Secondary structure	Descripción de la estructura secundaria	HELIX, SHEET
Connectivity	Anotación de la conectividad química	SSBOND, LINK, CISPEP
Miscellaneous	Características dentro de las de la macromolécula	SITE
Crystallographic	Descripción de la celda cristalográfica	CRYST1
Coordinate transformation	Operadores de transformación de coordenadas	ORIGXn, SCALEn, MTRIXn
Coordinate	Datos de coordenadas atómicas	MODEL, ATOM, ANISOU, TER, HETATM, ENDMDL
Connectivity	Resumen de conectividad química	CONNECT
Bookkeeping	Información de resumen	MASTER, END

Tabla 6.11. Estructura de un fichero PDB

6.5 OTRAS BASES DE DATOS GENÓMICAS

Son muchas y muy diversas las bases de datos genómicas existentes. No pretendemos, ni mucho menos, ofrecer un análisis exhaustivo pero sí dar al lector una guía de las principales.

6.5.1 Bases de datos de secuencias de ADN

Estas bases de datos almacenan información sobre secuencias de ADN que se han ido acumulando a lo largo de los años. Con el fin de garantizar la disponibilidad pública de secuencias, hay revistas que exigen como requisito previo a la publicación de un artículo el haber depositado la secuencia en alguna de las bases de datos que citaremos a continuación.

Existen tres grandes proyectos a nivel mundial:

- DDBJ (<http://www.ddbj.nig.ac.jp/>): el DDBJ (DNA Data Bank of Japan) depende del Instituto Nacional de Genética (NIG, National Institute of Genomics) y comenzó su andadura en 1986. Aunque el DDBJ recibe datos, principalmente, de investigadores japoneses, acepta contribuciones de otras nacionalidades.

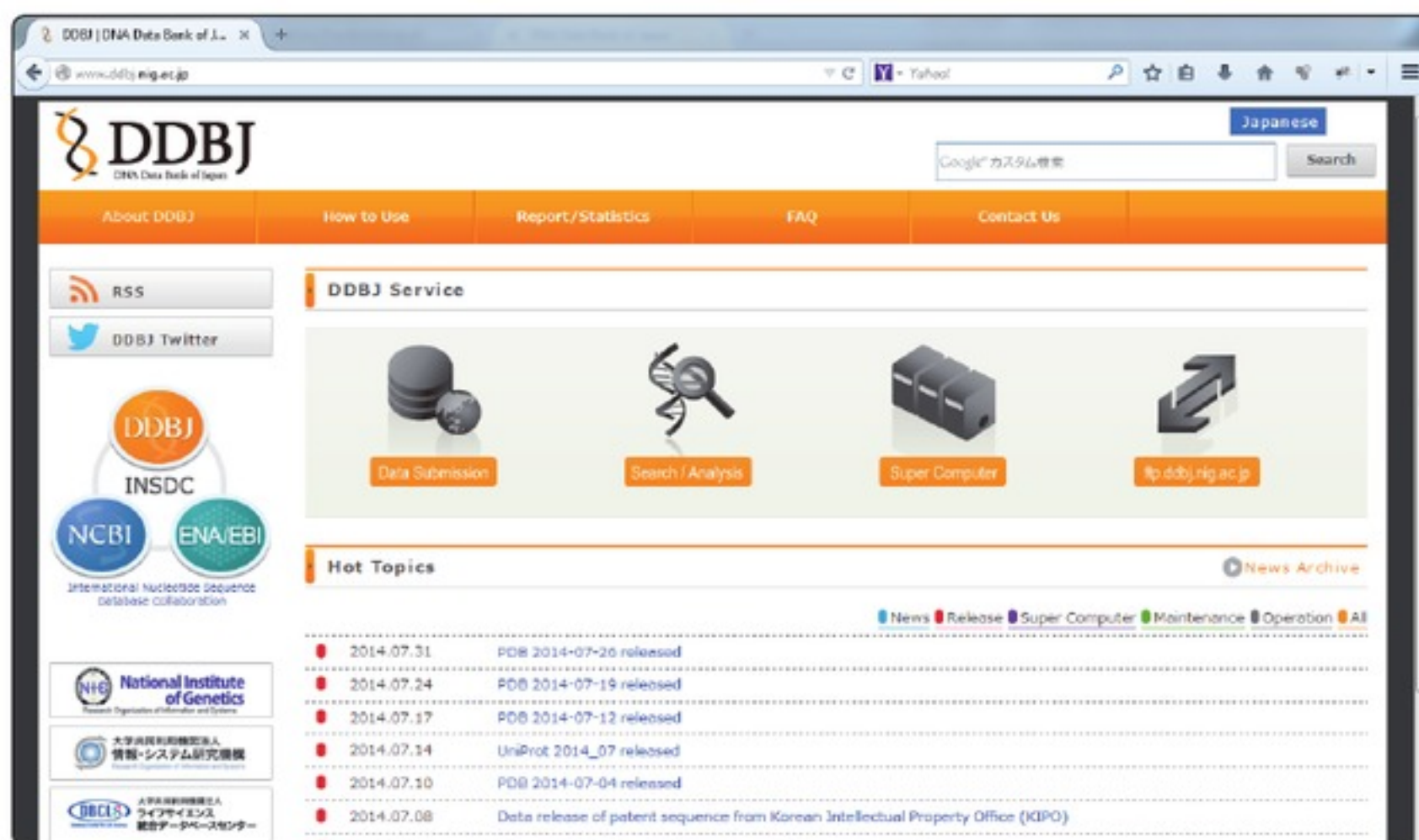


Figura 6.3. Página principal de DDBJ

- NCBI-GenBank** (<http://www.ncbi.nlm.nih.gov/genbank/>): es la base de datos de secuencias de ADN anotadas y públicas del NCBI. Se creó en 1982 por diversas organizaciones estadounidenses. Desde entonces ha ido creciendo a un ritmo vertiginoso, de hecho, se estima que su tamaño se duplica cada 18 meses.

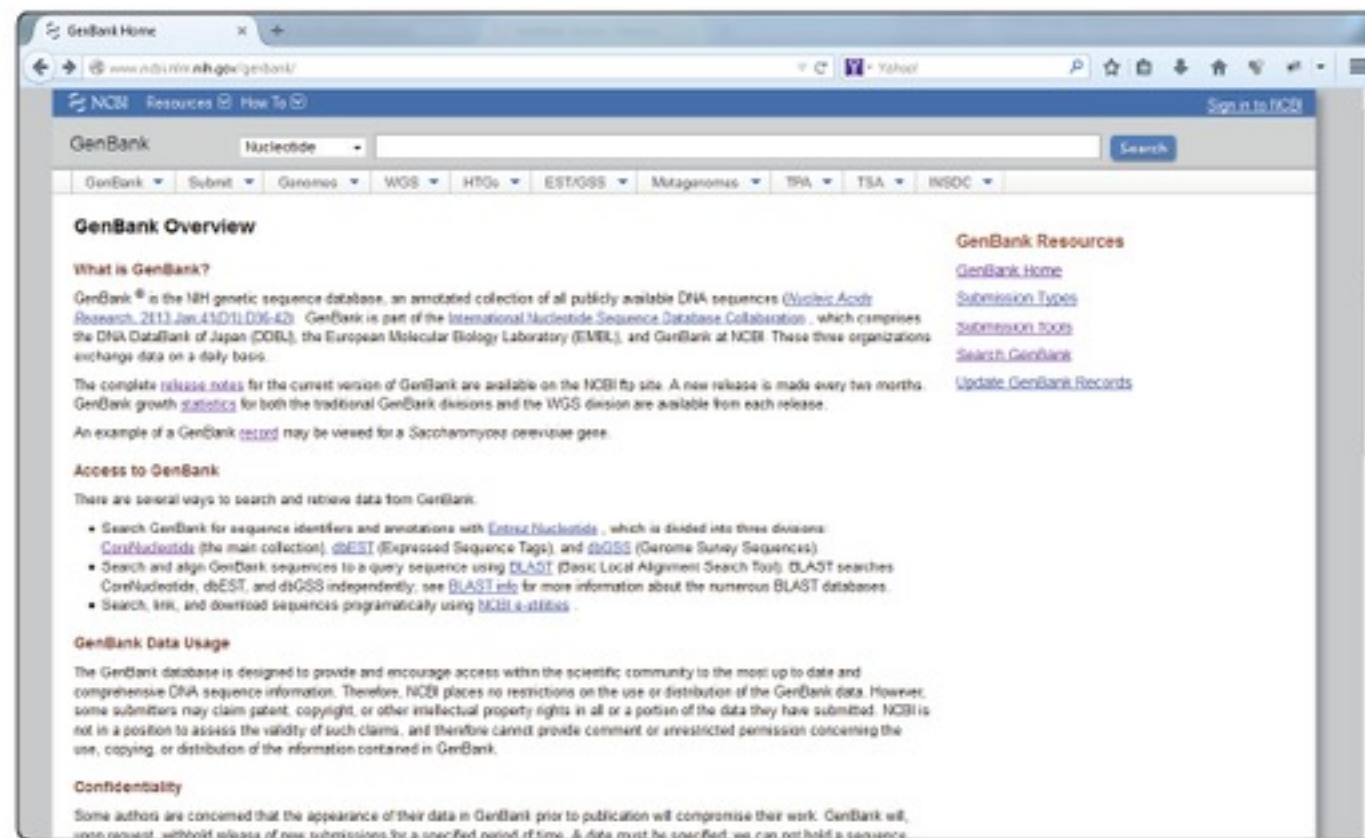


Figura 6.4. Página principal de GenBank

- EMBL-EBI** (<http://www.ebi.ac.uk/embl/index.html>): es un centro europeo de investigación y servicios bioinformáticos. Se constituyó en 1974, aunque se desarrolló durante la década de los 80.

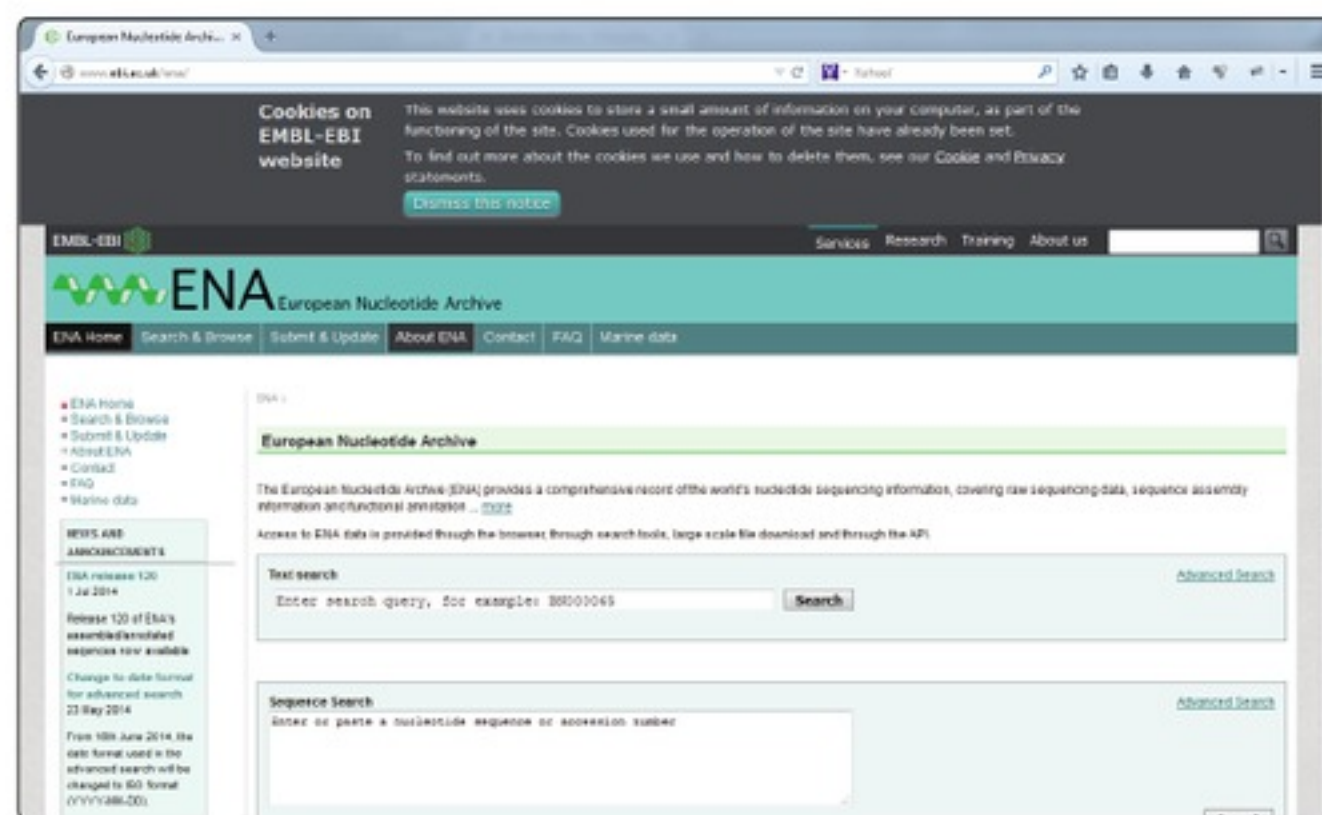


Figura 6.5. Página principal de EMBL-EBI

Estas tres instituciones comparten información, de manera que integran los datos de las investigaciones que tengan registradas y se sincronizan cada 24 horas.

6.5.2 Bases de datos de secuencias de ARN

Encontramos aquí dos bases de datos del NCBI. La primera de ellas, Unigen (<http://www.ncbi.nlm.nih.gov/unigene>), agrupa las secuencias almacenadas en GenBank de manera que cada grupo (también llamado *cluster*) contenga únicamente secuencias de un solo gen o gen putativo e información relacionada con el mismo.



Un gen putativo es un gen que originalmente no se encontraba en la secuencia de ADN sino que, debido a su utilidad práctica, se ha introducido por algún método artificial.

Los datos de Unigen se generan computacionalmente a partir de la información almacenada en otras bases de datos del NCBI, identificando los transcritos para el mismo locus y analizando su expresión cuando sea posible.

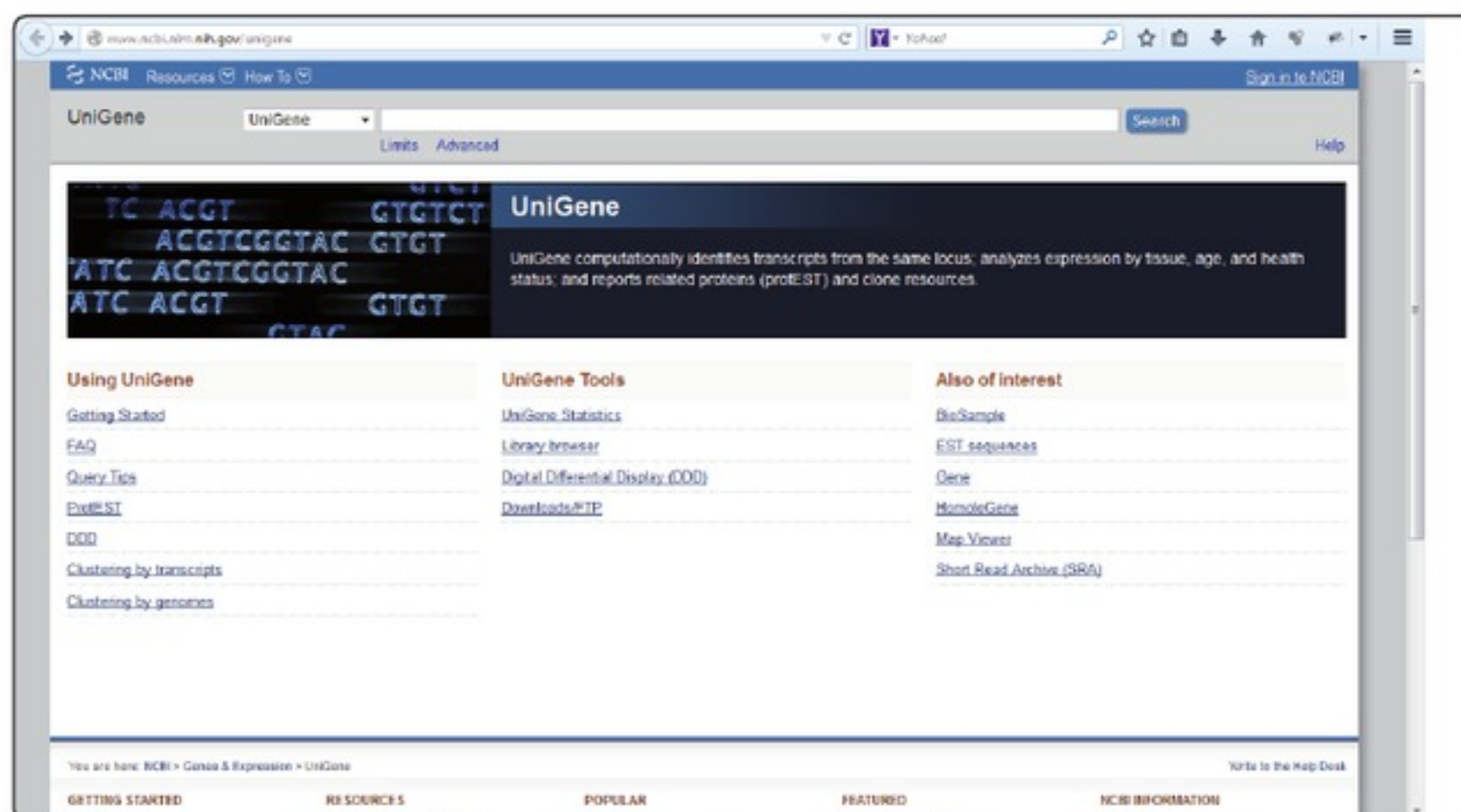


Figura 6.6. Página principal de NCBI-UniGene

La otra base de datos es dbEST (<http://www.ncbi.nlm.nih.gov/dbEST/index.html>). Se trata de una base de datos de EST (*Expressed Sequence Tags*) obtenida a partir de la información almacenada en GenBank.

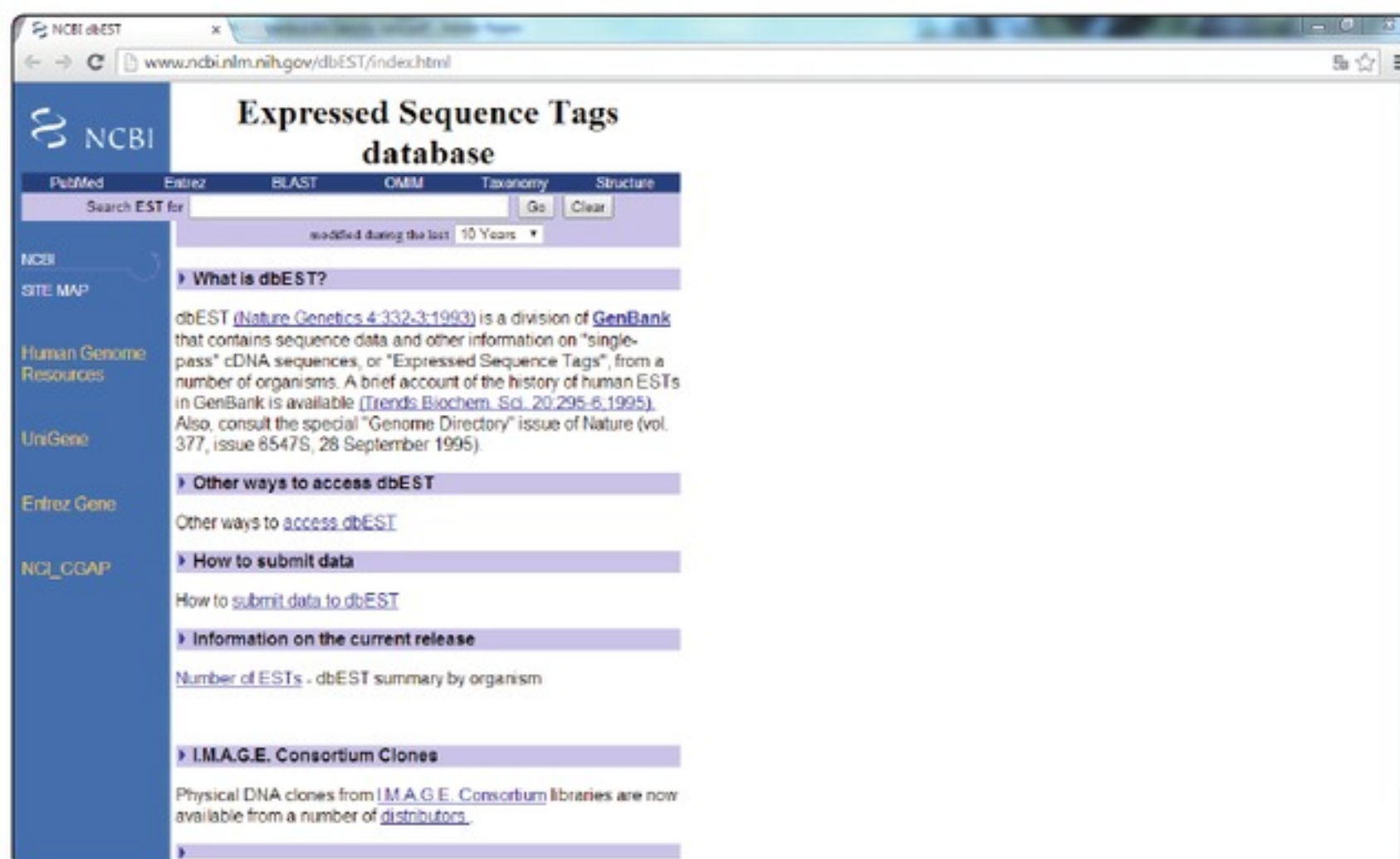


Figura 6.7. Página principal de dbEST

6.5.3 Bases de datos de secuencias de proteínas

Swiss-Prot (<http://www.expasy.ch/cgi-bin/sprot-search-ful>) es una base de datos curada manualmente que incluye referencias cruzadas en la que, a diferencia de otras, no existen registros duplicados.

El EBI dispone de TrEMBL (<http://www.ebi.ac.uk/trembl/>), que contiene secuencias de proteínas que todavía no se han incluido en Swiss-Prot, con la que mantiene un intercambio de información constante.

Finalmente, encontramos la base de datos PIR (<http://pir.georgetown.edu/>), mantenida entre el MIPS (Munich Information Centers for Protein Sequences) y el JIPID (Japanese International Protein Sequence Database). Almacena información sobre secuencias, anotaciones y alineamientos.

6.5.4 Bases de datos de patrones y perfiles

Estas bases de datos almacenan información sobre estructuras secundarias o dominios.

- Pfam (<http://pfam.xfam.org/>) es un proyecto de EMBL-EBI. Se trata de una base de datos de familias de proteínas con sus alineamientos múltiples de secuencia y sus modelos HMM (*Hidden Markov Models*).



En general, las proteínas están compuestas de una o más regiones funcionales denominadas dominios, de tal manera, que la combinación de dominios da lugar al amplio abanico de proteínas presentes en la naturaleza. La identificación de estos dominios dentro de la estructura de la proteína constituye un indicio de la función de la misma.

Dentro de PFAM existen dos componentes: PFAM-A, que son entradas curadas de alta calidad y que cubren un amplio rango de la base de datos de secuencias; y, PFAM-B, entradas generadas automáticamente a partir del procesado de la información contenida en la base de datos ADDA (*Automatic Domain Decomposition Algorithm*), accesible en http://ekhidna.biocenter.helsinki.fi/sqgraph/pairsdb/index_html. Aunque de menor calidad que la anterior, las familias PFAM-B son útiles para identificar regiones conservadas funcionalmente que no se encuentran en PFAM-A.



En la Práctica 3 se estudiará más en detalle la utilización de PFAM para encontrar dominios funcionales.

- PRODOM (<http://prodom.prabi.fr/>) es una base de datos de dominios de proteínas generados automáticamente mediante técnicas de agrupamiento de dominios de homólogos basándose en PSI-BLAST. Las secuencias de las proteínas origen se derivan de UnitProtKB (Swiss-Prot y TrEMBL). Cada entrada de PRODOM proporciona un alineamiento múltiple con los dominios de homólogos y una secuencia de consenso de la familia.

- PRINTS (<http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php>) es un archivo genómico de huellas digitales de proteínas. La huella digital de una proteína es un grupo de motivos conservados empleado para caracterizar dicha proteína.
- BLOCKS (<http://blocks.fhcrc.org/>) es una base de datos de regiones conservadas de grupos de proteínas documentados en InterPro (<http://www.ebi.ac.uk/interpro/>). La información contenida en BLOCK se genera automáticamente.
- La base de datos PROSITE (<http://prosite.expasy.org/>) es una base de datos de familias de proteínas y dominios. Se basa en la existencia de un gran número de proteínas que pueden agruparse siguiendo criterios como la similitud de sus secuencias, en un número limitado de familias, de tal manera, que las proteínas o dominios de proteínas pertenecientes a una misma familia suelen tener atributos funcionales comunes y proceden de un ancestro común.

6.5.5 Bases de datos clínico-genéticas

- OMIM (<http://omim.org/>) es un catálogo de información sobre el ser humano que describe informaciones genéticas y relaciones genotipo-fenotipo.
- HGMD (<http://www.hgmd.org/>) es una recopilación de mutaciones genéticas conocidas y publicadas del ser humano.
- GeneCards (<http://www.genecards.org/>) es una base de datos secundaria que extrae e integra información genómica, transcriptómica, proteómica, genética, clínica y funcional de diversas fuentes de datos sobre genes humanos. Entre la información que se incluye se encuentra relación con enfermedades, mutaciones, SNP, expresión y función génica, etc.
- NCBI-PheGenI (<http://www.ncbi.nlm.nih.gov/gap/phegeni/>) ofrece relaciones entre las variaciones de las secuencias de ADN y diferencias entre genes y la expresión de los mismos para un fenotipo dado, como el asma o la diabetes. Los resultados se enlazan a las fuentes originales en donde los investigadores pueden encontrar mayor información.

6.5.6 Bases de datos de mutaciones y SNP

Un aspecto crítico en el análisis genético es la relación entre genotipo y fenotipo y cómo una variación en el genotipo tiene una manifestación apreciable en un rasgo hereditario (fenotipo).

- Los SNP (*Single Nucleotide Polymorphism*) son las variaciones más comunes y constituyen un cambio en un único nucleótido.
- dbSNP (<http://www.ncbi.nlm.nih.gov/snp>) es una base de datos pública de polimorfismos, ya sean sustituciones de un solo nucleótido o pequeñas inserciones o borrados de varias bases.
- GWAS Central (<http://www.gwascentral.org/>) es el repositorio central de mutaciones de la Human Genome Variation Society (HGVS). Es una base de datos curada.

6.5.7 Bases de datos de genómica funcional

La Genómica Funcional es la rama de la Genómica que se encarga del estudio de las relaciones genotipo-fenotipo, es decir, de cómo los genes determinan las características de los individuos a los que pertenecen.

- ArrayExpress (<http://www.ebi.ac.uk/arrayexpress/>) es un repositorio público de experimentos de genómica funcional del que es posible consultar y descargar información. Incluye información de expresión génica procedente de estudios realizados con microarrays y técnicas de secuenciación de alto rendimiento.
- GEO (<http://www.ncbi.nlm.nih.gov/geo/>) es un repositorio de datos público que almacena y distribuye gratuitamente información sobre microarrays y otros datos genómicos.
- SMD (<http://smd.princeton.edu/>) almacena los datos experimentales de microarrays sin procesar y normalizados, así como sus correspondientes archivos de imagen. Además, proporciona interfaces para la recuperación de datos, análisis y visualización. Los datos son hechos públicos según criterio del investigador o en el momento de su publicación.

PRÁCTICA 2: BÚSQUEDA DE SECUENCIAS

La búsqueda y recuperación de secuencias de interés de distintas bases de datos es una de las tareas más comunes en Bioinformática. Básicamente, se trata de recuperar una secuencia de nucleótidos o de aminoácidos de una base de datos biológica. Sin embargo, pese a la aparente sencillez de esta tarea, realizarla de una manera eficiente y eficaz requiere cierta práctica.

En esta práctica se estudiarán con cierto nivel de detalle varios ejemplos de búsqueda de secuencias en distintas bases de datos biológicas. No se pretende, por razones de extensión, cubrir absolutamente todas las bases de datos biológicas ni tampoco todas las posibilidades de análisis existentes pero sí hacer hincapié en los aspectos más importantes a tener en cuenta y en la interpretación de los resultados. Por esta razón, para una explicación más amplia sobre la configuración de las búsquedas avanzadas, remitimos al lector a la Bibliografía y a la documentación existente sobre cada base de datos.

7.1 SECUENCIAS DE ORGANISMOS PROCARIOTAS

En los organismos procariotas, el tamaño limitado de sus genes, así como el hecho de que la relación entre la secuencia de ADN y el ARNm sea lineal, tiene como principal consecuencia que la información almacenada en las bases de datos sea fácil de entender y de anotar. En esta primera práctica, estudiaremos la información que ofrece GenBank sobre el gen dUTPase de la *Escherichia Coli*.



NCBI-Entrez permite consultar múltiples bases de datos simultáneamente. Sin embargo, las búsquedas están restringidas siempre a las diferentes bases de datos del NCBI.

Para acceder a Entrez, abriremos una ventana del navegador y teclearemos la dirección de su página principal (<http://www.ncbi.nlm.nih.gov/entrez/>). Puesto que vamos a buscar secuencias de ADN, seleccionaremos, en el menú desplegable de la parte superior izquierda, la opción **Nucleotide** (ver Figura 7.1).

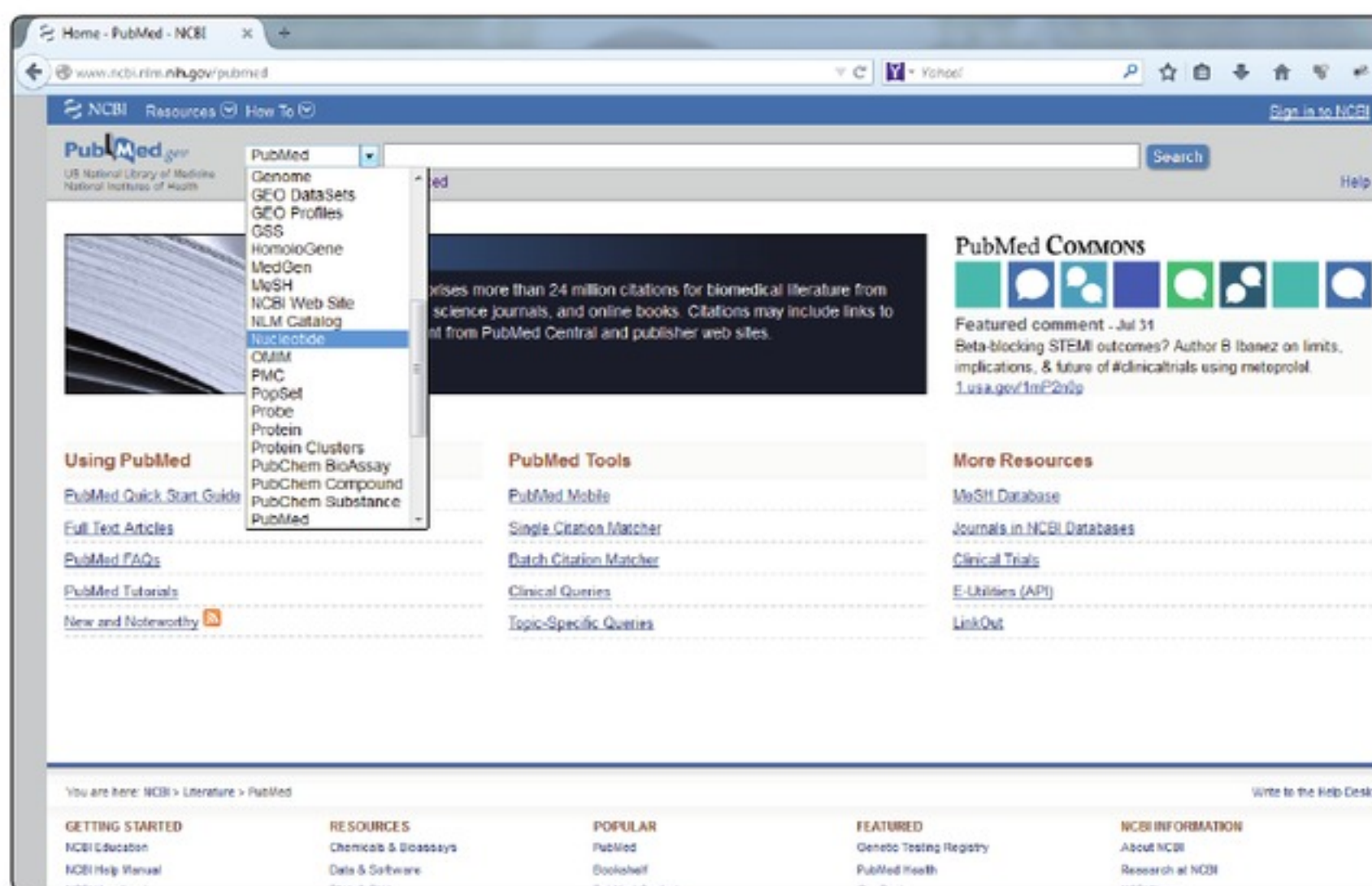


Figura 7.1. Página principal de búsqueda del NCBI

Por supuesto, es posible escoger cualquiera de las otras opciones y la búsqueda se efectuaría sobre otras bases de datos del NCBI. De la misma manera, si se selecciona la opción **All Databases**, se obtendrá una perspectiva general de toda la información almacenada en el NCBI acerca de la secuencia introducida (ver Figura 7.2).

The screenshot shows the NCBI Entrez search engine interface. At the top, there's a navigation bar with links to Home, Search, Site Map, PubMed, All Databases, Human Genome, GenBank, Map Viewer, and BLAST. Below this is a search bar with the text 'Search a cross databases' and a search box containing 'X01714'. To the right of the search box are buttons for 'Go', 'Submit', and 'Help'.

The search results are displayed in a grid format, showing the number of results found in each database. The results are as follows:

Database	Results
PubMed: biomedical literature citations and abstracts	50871
PubMed Central: free, full text journal articles	13510
Site Search: NCBI web and FTP sites	45
Books: online books	499
OMIM: online Mendelian inheritance in Man	126
OMIA: Online Mendelian inheritance in Animals	none
Nucleotide: sequence database (GenBank)	167470
Protein: sequence database	168532
Genome: whole genome sequences	5
Structure: three-dimensional macromolecular structures	680
Taxonomy: organisms in GenBank	1
SNP: single nucleotide polymorphism	7742
Gene: gene-centered information	1098
HomoloGene: eukaryotic homology groups	469
PubChem Compound: unique small molecule chemical structures	7
PubChem Substance: deposited chemical substance records	833
Genome Projects: genome project information	2
UniGene: gene-oriented clusters of transcript sequences	145
CDD: conserved protein domain database	6
3D Domains: domains from Entrez Structure	2624
UniSTS: markers and mapping data	130
PopSet: population study data sets	1813
GEO Profiles: expression and molecular abundance profiles	5561
GEO DataSets: experimental sets of GEO data	13
Cancer Chromosomes: cytogenetic databases	1
PubChem BioAssay: bioactivity screens of chemical substances	126
GENSAT: gene expression atlas of mouse central nervous system	8088
Probe: sequence-specific reagents	251
Journals: detailed information about the journals indexed in PubMed and other Entrez databases	none
NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections	237
MeSH: detailed information about NLM's controlled vocabulary	66

At the bottom of the results, there is a note: "Result counts displayed in gray indicate one or more terms not found".

Figura 7.2. Selección de bases de datos de búsqueda



Vuelve a la página principal de búsqueda y selecciona alguna de las opciones que no sea **Nucleotide** y comprueba la información que te ofrece el NCBI sobre la secuencia bajo estudio desde distintas perspectivas.

En la caja de texto de búsqueda, teclee el identificador del gen que buscamos, X01714 y haga clic en **Go**.

La Figura 7.3 muestra la entrada de GenBank X01714. Aunque por defecto la información se muestra en formato GenBank, el menú desplegable de la parte superior izquierda permite indicar también otros formatos de salida, el más común de los cuales es el FASTA.

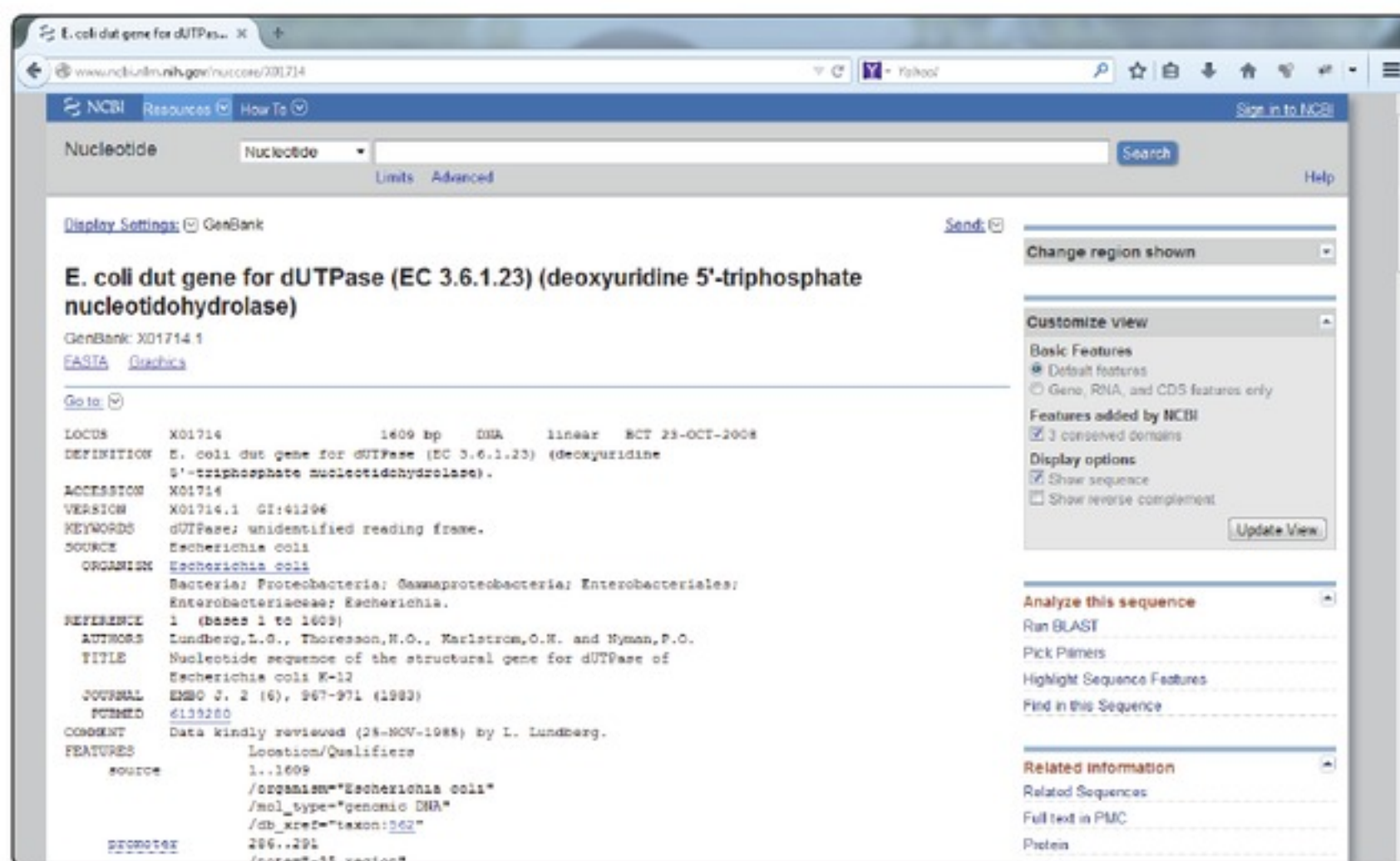


Figura 7.3. Resultados de la búsqueda

Ahora cambiamos la vista de **GenBank** a **Summary** y en la página de resultados (ver Figura 7.4) aparecerá una definición breve de la secuencia de entrada, precedida de su identificador.

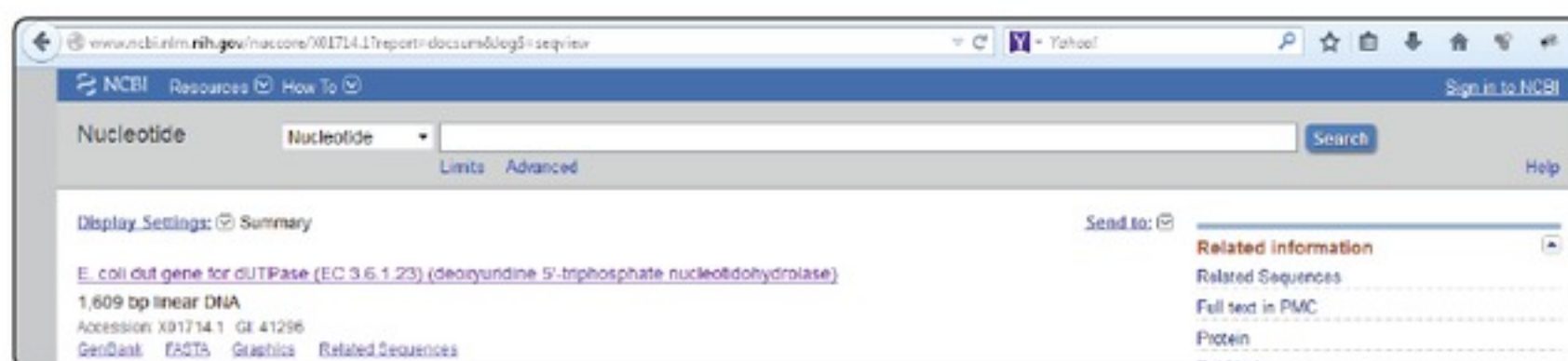


Figura 7.4. Registro GenBank X01714

Para guardar la secuencia en un fichero de texto en formato FASTA, selecciona la opción de formato texto (**Text**) y **Send To** en el menú desplegable para generar el fichero de la entrada y guárdalo utilizando las opciones del navegador.


```

>X01714.1|emb|X01714.1| E. coli dut gene for dUTPase (EC 3.6.1.23) (deoxyuridine
5'-triphosphate nucleotidohydrolase)
CAGAGAAATCAAAAGCAGGCCAGCGGGTGATGAATTAACAATAAAATGGTTAAAAACCCGATAT
CGTCGACAGGCGTTGCCGCACTAAAGACCATCGACCCCTACGTCGTTGGATTGGCCGCAACAAATAAT
GTGGAGAAATACGCCCGCAAAACGATCCGTAAAAACCTTGATCTGATCTGCCGCAACGATGTTCCC
AGCCAACTCAAGGATTTAAGCAGCAGCAACACCGATTACACCTTTCTGGCAGGACGAGATAAGTCTT
ACCCCTTGAGCGCAAGAGCTCCTTGGCCAAATTATTCTCGACGAGATCGTGACCCGTTATGATGAAAA
AATCGAGCGTTAAGATTCTGGACCCGCGCGTTGGGAGGAAATTCGGCTCCCGAGTTATGCCACCTCTGGC
TCTGCCGAGCTTGACCTGCGTGGCTGTCTCAACGACGCGCTAGAACTGGCTCCGGGTGACACTACGCTGG
TTCCGACCCGCGCTGGCGATTCTATTTGCCGATCCTTCACTGCCGCAATGATGCTGCCGCGCTCCGATT
GGGACATAAGCAGCTATCCTGCTTGGTAACCTGGTAGGATTGATCGATTCTGACTATCAGGCGCACTG
ATGATTTCCGTTGGGAAACCGTGGTCAGGACAGCTTCAACATTCAACCTGGCGAAGCATCGCCAGATGA
TTTTGTTCCGGTAGTACAGGCTGAATTTAATCTGGTGGAGATTTCGACGCCACCGACCGCGGTGAAGG
CGGCTTTGGTCACTCTGGTCTGATACACATACGCTCCGATACAGCTATACATAGCCGCAACAT
TTCTGTTTGGGCTATAGCGTGGGTGCCGCTGGCAAGTGCTTATTTTCAGGGGATTTTGTAACTGCA
GAAACCAAACTCGGAAAGGACCGCTCCGAGGAAATACTTCACTCTCTGCCGCTGATGCTGGATCCA
CGGATGGAGCGCAACGATCACGACGGCAAACTGGCGCGCTCTGTCGCGCTTTCGGAAGCGGCACTGTA
TCGACACTTCCCACTAAGACCCGATGTTCTATAGCTGATTCGATTCGATTCGATTCGATTCGATTCGATTC
CGCATCAACCTGATTCTGAAAGATGAAAGACCAACAGCGCGCTGCTGATTGTTGTTGCTGCTTC
TCGGTTTTGGTGAAGCTAATCTGGCTGACCCGATCCTCACTGGTCATGCGCTAATGTTTGAACAGGA

```

Figura 7.5. Registro GenBank X01714 en formato texto

A continuación, aprenderemos a interpretar el contenido del fichero, aunque para una descripción más detallada y formal remitimos al lector al capítulo anterior, en donde se trató el formato de fichero de GenBank.

La Tabla 7.1 resume la información que se extrae de la cabecera:

Nombre del campo	Interpretación
LOCUS	Nombre del locus (ECDUT), tamaño de la secuencia de nucleótidos (bp), naturaleza de la molécula (ADN) y topología (lineal)
DEFINITION	Descripción corta del gen al que corresponde la secuencia de esta entrada. En este caso, el gen DUTPase de la <i>E. Coli</i>
ACCESSION	Identificador de la secuencia (X01714)
VERSION	Versión de la secuencia
KEYWORDS	Lista de términos que caracterizan la entrada
SOURCE	Nombre común del organismo al que pertenece la secuencia
ORGANISM	Información taxonómica
REFERENCE	Información bibliográfica de toda la secuencia o solamente de partes de la misma
COMMENT	Texto en formato libre que proporciona información adicional al resto de campos

Tabla 7.1. Información de la cabecera del fichero

Después de la cabecera, encontramos la sección de características (**FEATURES**) que describe las regiones del gen y las propiedades biológicas que se han identificado en la secuencia de nucleótidos. Llegados a este punto, conviene hacer hincapié en la simplicidad de la transmisión de la información genética en este tipo de organismos (ver Figura 7.6). Como puede verse, la secuencia de ARNm que se traduce en proteína es aquella que está señalizada con una subsecuencia especial llamada RBS (*Ribosome Binding Site*).

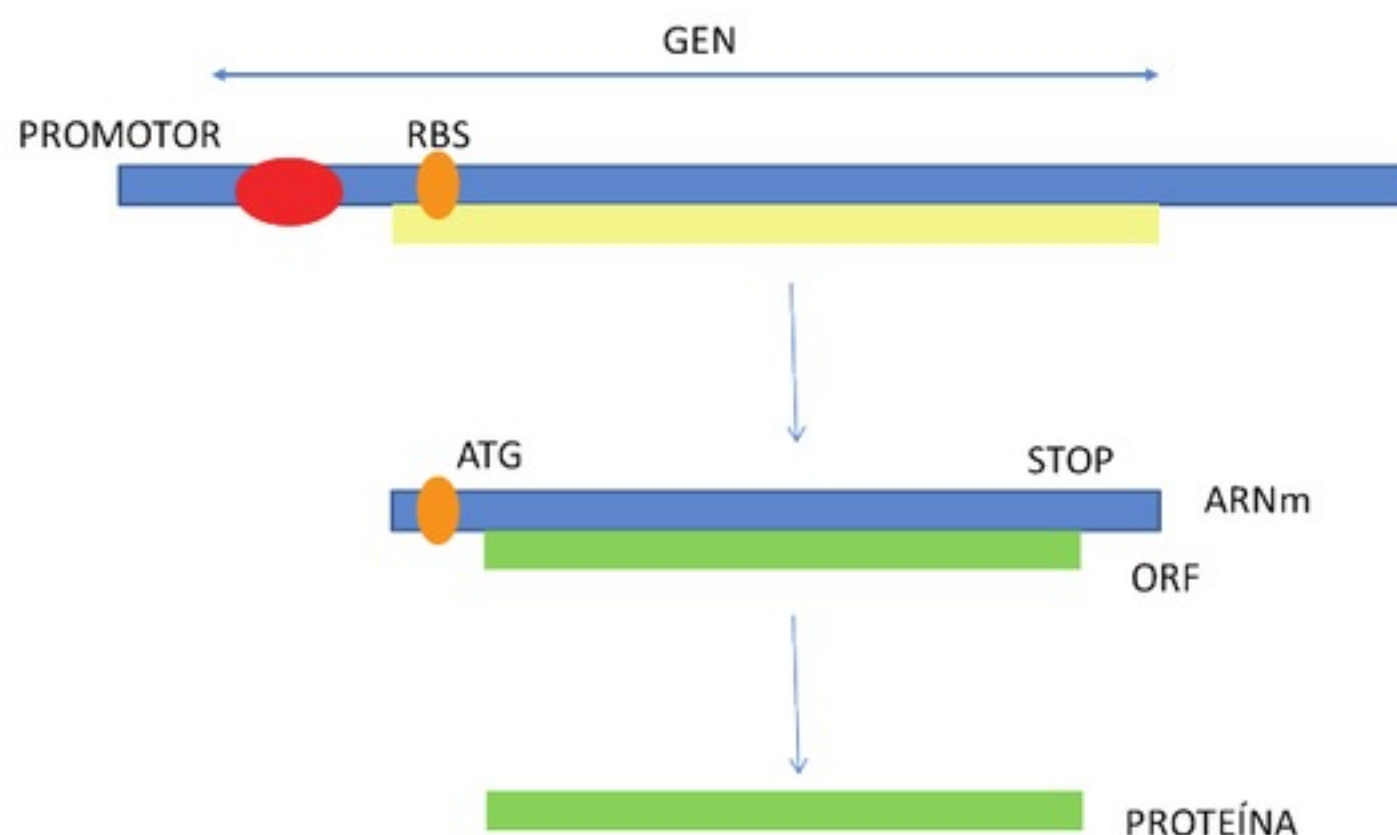


Figura 7.6. Relación entre gen, ARNm y secuencia de proteínas en las células procariotas

De acuerdo con lo anterior, la información registrada en la base de datos debe contener, al menos, las coordenadas de algún promotor, las coordenadas del RBS y las coordenadas de los límites de la región ORF. El resumen de la información de la entrada obtenida está recogido en la Tabla 7.2:

Campo	Significado
/source	Origen de la zonas específicas de la secuencia. Es muy útil cuando se necesita distinguir vectores de clonación de secuencias huésped. En el caso del X01714, la secuencia completa corresponde al ADN genómico de la <i>E. Coli</i>
/promotor	Coordenadas del promotor. En este caso, hay dos promotores: uno en la región -35 (posiciones 286 a 291) y otro en la región -10 (posiciones 310 a 316)
/misc_feature	Ubicación putativa del comienzo de la transcripción (síntesis del ARNm). En este caso, es la secuencia comprendida entre las posiciones 322 a 324
RBS	Localización del último elemento de subida, que se encuentra en las posiciones 330 a 333
CDS	ORF del gen. La primera línea son las coordenadas de la ORF, desde su codón inicial hasta el final, esto es, posiciones 343 a 798. El resto de líneas indican el código genético que hay que aplicar (<i>/transl_table</i>), el identificador de la proteína (<i>/protein_id</i>) y referencias en otras bases de datos (<i>/db_xref</i>). Finalmente, <i>/translation</i> es la secuencia de aminoácidos del segmento CDS

Tabla 7.2. Sección FEATURES de la entrada de GenBank



Las entradas con más de un gen son muy frecuentes en los registros de GenBank.

7.2 SECUENCIAS DE ORGANISMOS EUCARIOTAS

En los organismos eucariotas las regiones codificantes de ADN se dividen en un número variable de exones (fragmentos del gen que contribuyen a la proteína final) entrelazados con intrones (fragmentos del gen que no codifican).

A continuación, se analizará la entrada de GenBank V00505, que contiene el gen que codifica la delta-globina humana. Para ello, abrimos un navegador web y tecleamos la dirección de la página principal del NCBI: <http://www.ncbi.nlm.nih.gov/entrez/> y, al igual que hemos hecho en el epígrafe anterior, en el menú desplegable de la parte superior izquierda, seleccionamos la opción **Nucleotide**.

En la caja de texto de búsqueda, tecleamos el identificador del gen que buscamos, V00505, hacemos clic en **Go** y obtenemos el siguiente registro:

```

LOCUS      V00505              1976 bp  DNA   linear  PRI 14-NOV-2006
DEFINITION Human gene for delta-globin.
ACCESSION  V00505
VERSION    V00505.1 GI:30510
KEYWORDS   delta globin; germ line; globin.
SOURCE     Homo sapiens (human)
ORGANISM   Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates;
            Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE  1 (bases 1 to 1976)
AUTHORS    Spritz,R.A., DeRiel,J.K., Forget,B.G. and Weissman,S.M.
TITLE      Complete nucleotide sequence of the human delta-globin
gene
JOURNAL    Cell 21 (3), 639-646 (1980)
PUBMED     7438204
COMMENT    KST HSA.DELGLOBIN.
FEATURES   Location/Qualifiers
  
```

```

source      1..1976
             /organism="Homo sapiens"
             /mol_type="genomic DNA"
             /db_xref="taxon:9606"
prim_transcript 123..1763
exon        123..265
             /number=1
CDS         join(173..265,394..615,1505..1633)
             /codon_start=1
             /product="delta globin"
             /protein_id="CAA23763.1"
             /db_xref="GI:30511"
             /db_xref="GDB:119298"
             /db_xref="GOA:P02042"
             /db_xref="HGNC:4829"
             /db_xref="InterPro:IPR000971"
             /db_xref="InterPro:IPR002337"
             /db_xref="InterPro:IPR009050"
             /db_xref="InterPro:IPR012292"
             /db_xref="PDB:1SHR"
             /db_xref="PDB:1SI4"
             /db_xref="UniProtKB/Swiss-Prot:P02042"
             /translation="MVHLTPEEKTAVNALWGKVNVDAGVGEALGRLLVVYPWTQ
RFFESFGDLSASPDVAMGNPKVKAHGKKVLGAFSDGLAHLNKLKGTFSQLSSELHCDKLHVDPENF
RLLGNVLCVLRNFGKEFTTPQMQAAYQKVVAGVANALAHKYH"
intron      266..393
             /number=1
exon        394..615
             /number=2
intron      616..1504
             /number=3
exon        1505..1763
             /number=3

ORIGIN
    1 aatgaagggt catttttcat ttcacaaac taatgaaacc ctgcttatct
    taaaccaacc
    61 tgctcactgg agcagggagg acaggaccag cataaaaggc agggcagagt
    cgactgttgc
    121 ttacactttc ttctgacata acagtgttca ctagcaacct caaacagaca
    ccatggtgca
    181 tctgactcct gaggagaaga ctgctgtcaa tgccctgtgg ggcaaagtga
    acgtggatgc
    241 agttggtggt gaggccctgg gcaggttggt atcaaggtta taagagaggc
    tcaaggaggc
    301 aaatggaaac tgggcatgtg tagacagaga agactcttgg gtttctgata
    ggcaactgact
    361 ctctgtccct tgggctgttt tctaccctc agattactgg tggctctaccc

```



```
ttggacccag
  421 aggttctttg agtcctttgg ggatctgtcc tctcctgatg ctgttatggg
caaccctaag
  481 gtgaaggctc atggcaagaa ggtgctaggt gccttttagtg atggcctggc
tcacctggac
  541 aacctcaagg gcactttttc tcagctgagt gagctgcact gtgacaagct
gcacgtggat
  601 cctgagaact tcaggggtgag tccaggagat gcttcacttt tctcttttta
ctttctaatac
  661 ttacattttg gttctttttac ctacctgctc ttctcccaca tttttgtcat
tttactatat
  721 tttatcattt aatgcttcta aaattttggt atttttttat ttaaaaattc
tgcatttttt
  781 ccttcctcac aatcttgcta ctctaaatta ttaatatcc tgtctttctc
tcccaacccc
  841 ctcccttcat ttttccttct ctaacaacaa ctcaaattat gcataccagc
tctcacctgc
  901 taatttgcga cttagaataa tccttttggt tctccacatg ggtatgggag
aggctccaac
  961 tcaaagatga gaggcataga atactgtttt agaggctata aatcatttta
caataaggaa
 1021 taattggaat tttataaatt ctgtagtaaa tggaatggaa aggaaagtga
atatttgatt
 1081 atgaaagact aggcagttac actggagggtg gggcagaagt cgttgctagg
agacagccca
 1141 tcatcacact gatttatcaa ttcaatttgt atctattaat ctgtttatag
taattaattt
 1201 gtatatgcta tatacacata caaaattaaa actaatttgg aattaatttg
tatatagtat
 1261 tatacagcat atatgtacat atatagacta catgctagtt aagtacatag
aggatgtgtg
 1321 tgtatagata tatgttatat gtatgcattc atatatgtac ttatttatgc
tgatgggaat
 1381 aacctgggga tcagttttgt ctaagatttg ggcagaaaaa aatgggtggt
ggctcagttc
 1441 tcagaagcca gtctttattt ctctgttaac catatgcatg tatctgcta
cctcttctcc
 1501 gcagctcttg ggcaatgtgc tgggtgtgtg gctggcccgc aactttggca
aggaattcac
 1561 cccacaaatg caggctgcct atcagaaggt ggtggctggt gtggctaata
ccttggtca
 1621 caagtaccat tgagatcctg gactgtttcc tgataaccat aagaagacc
tatttccta
 1681 gattctattt tctgaacttg ggaacacaat gcctacttca aggttatggc
ttctgctaa
 1741 taaagaatgt tcagctcaac ttctgatta atttactta tttcattttt
ttgtccaggt
```

```

1801 gtgtaagaag gttcctgagg ctctacagat agggagcact tctttatattt
acaaagagta
1861 catgggaaaa gagaaaagca aggaaccgt acaaggcatt aatgggtgac
acttctacct
1921 ccaaagagca gaaattatca agaactcttg atacaaagat aatactggca
ctgcag
//

```

Figura 7.7. Registro GenBank correspondiente a la entrada V00505

Este gen tiene dos intrones (el primero entre las posiciones 266 y 293 y el segundo entre la 616 y la 1504) y dos exones (el primero entre las posiciones 394 y 615 y el segundo entre la 1505 y la 1763), como marcan los campos /exon y /intron.

7.3 BÚSQUEDA DE VARIACIONES

Encuentre el efecto de una variación en la región 3713 del gen humano que determina la distrofia muscular de Duchenne (DMD AND Homo sapiens).

Para resolver este análisis, abrimos un navegador web y tecleamos la dirección de la página principal de Entrez. De nuevo, en el menú desplegable de la parte superior izquierda, seleccionamos la opción **Nucleotide**. En la caja de texto de búsqueda, escribimos “DMD AND Homo sapiens” y hacemos clic en **Go**.

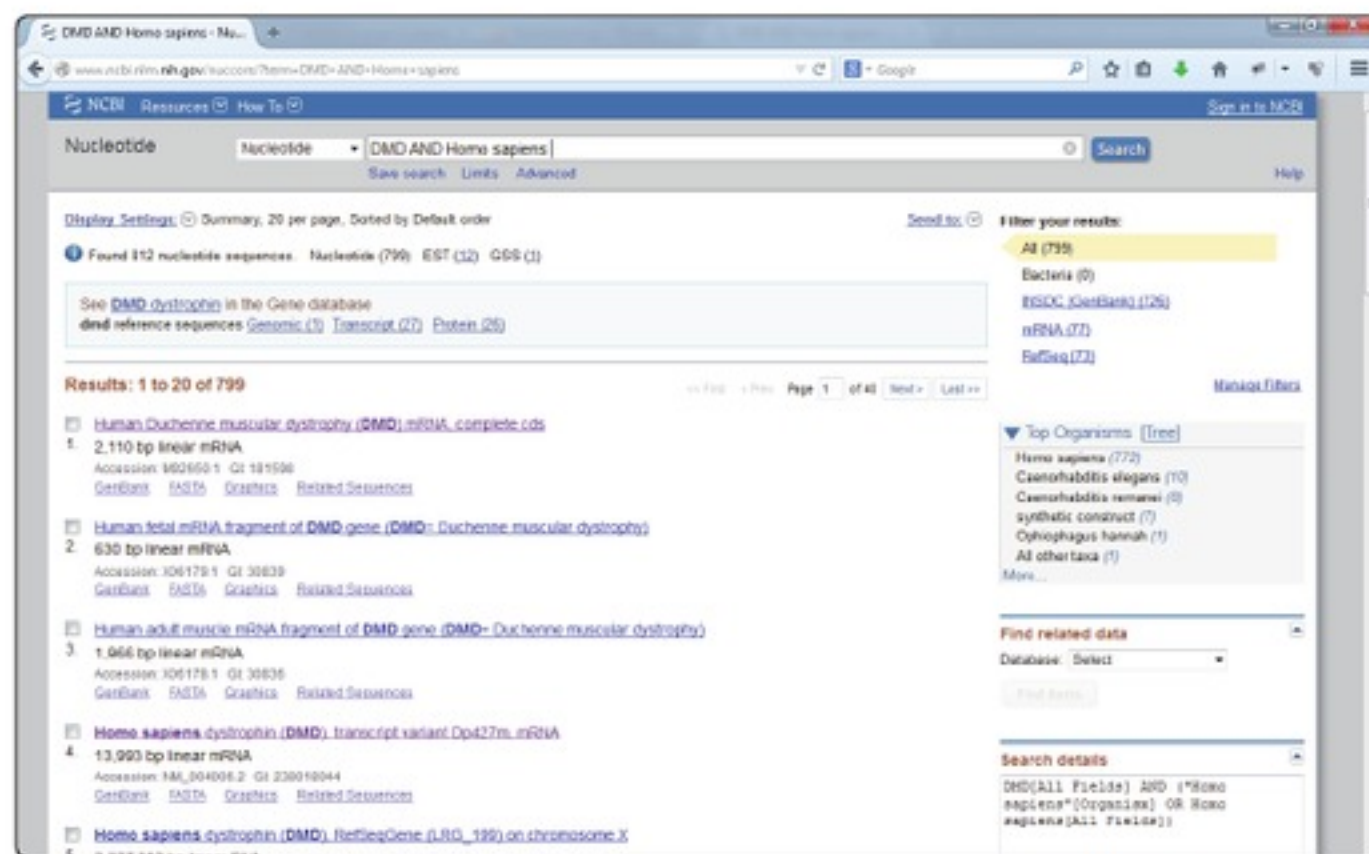


Figura 7.8. Búsqueda del registro en GenBank

En la parte superior derecha, seleccionamos **RefSeq** y el transcrito Dp427m (*accession* NM_004006).

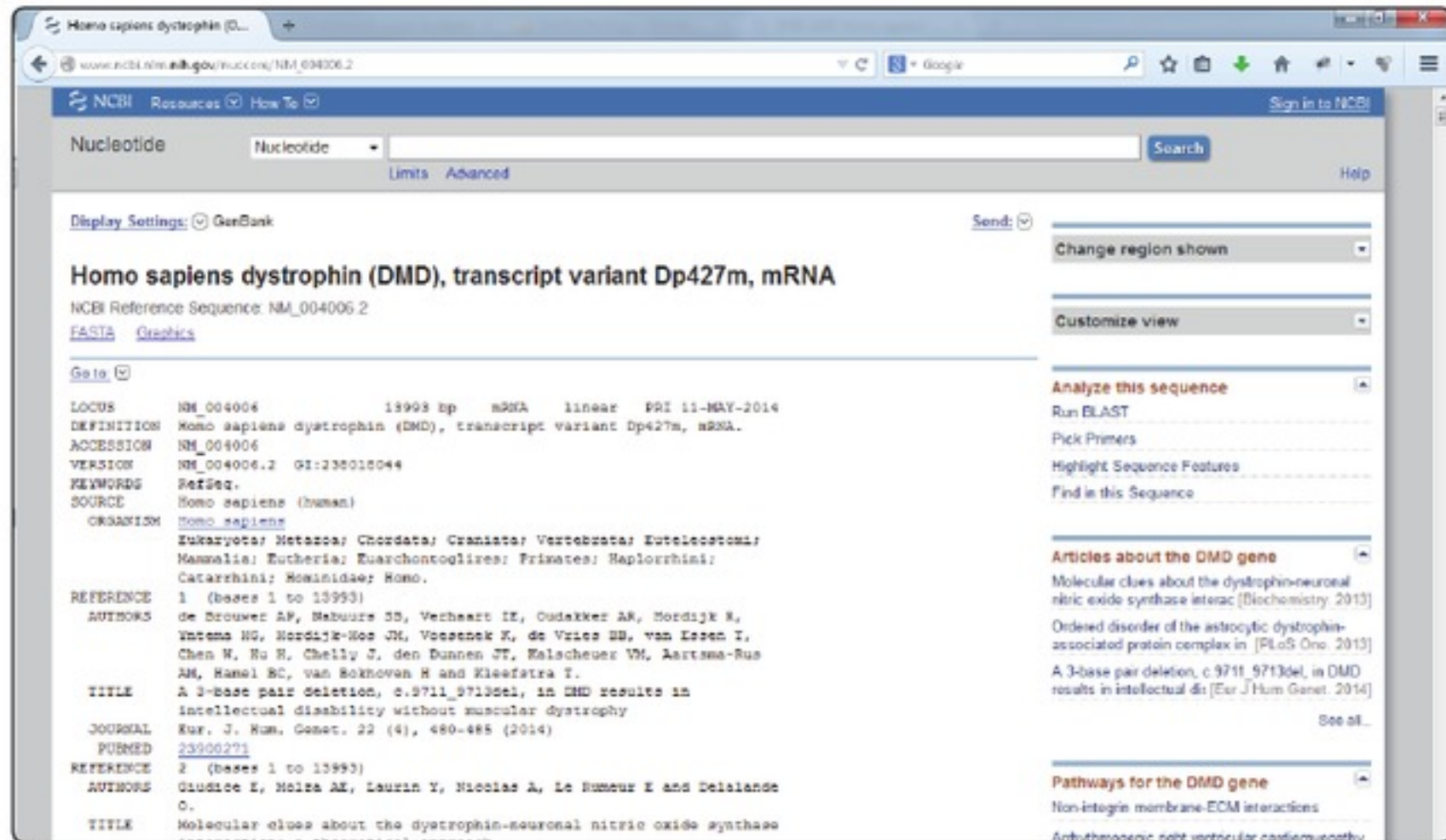


Figura 7.9. Registro de GenBank

Si examinamos, como se nos pide, la sección del fichero correspondiente a la variación 3713, encontramos:

```
variation      3713
               /gene="DMD"
               /gene_synonym="BMD; CMD3B; DXS142; DXS164; DXS206;
DXS230;
               DXS239; DXS268; DXS269; DXS270; DXS272; MRX85"
               /note="point mutation causing translational stop;
Glu1157X
               (S43366)"
               /phenotype="Duchenne Muscular Dystrophy (DMD)"
               /replace="t"
```

Figura 7.10. Sección del fichero correspondiente a la variación 3713

Es decir, se trata de una variación que termina el proceso de traducción del gen.

7.4 EJEMPLO DE ESTUDIO DE UNA PROTEÍNA

Se trata de realizar un estudio práctico de la proteína humana hemoglobina subunidad gamma-1 codificada en el HBG1, respondiendo a las siguientes preguntas:

1. ¿Cuáles son las coordenadas del gen HBG1?
2. ¿En qué hebra se encuentra el gen HBG1?
3. ¿Cuántos transcritos codifica el gen HBG1?

Como viene siendo habitual, abrimos un navegador web y tecleamos la dirección de Entrez (<http://www.ncbi.nlm.nih.gov/entrez/>). En el menú desplegable de la parte superior izquierda, seleccionamos la opción **Nucleotide** y en la caja de texto de búsqueda, tecleamos el identificador del gen que buscamos, HBG1 y hacemos clic en **Go**. En los resultados de la búsqueda (ver Figura 7.11), seleccionamos la séptima entrada por ser la que contiene toda la región codificante.

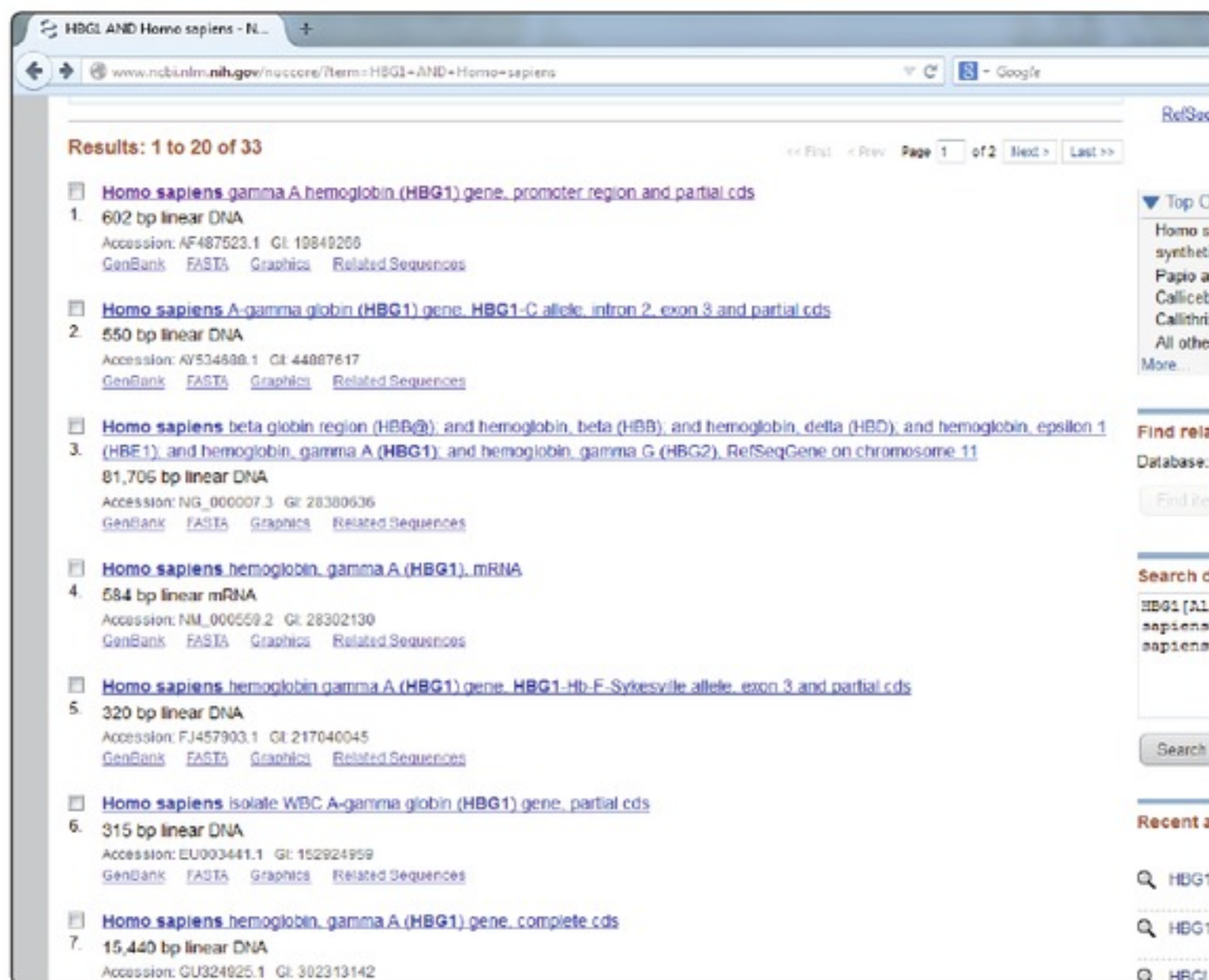


Figura 7.11. Resultados de la búsqueda en Entrez

Si visualizamos el fichero en formato GenBank, veremos que la información correspondiente al gen HBG1 es:

```
gene      2006..3591
          /gene="HBG1"
```

Y que el ARNm que codifica la proteína es:

```
mRNA      join(2006..2150,2273..2495,3376..3591)
          /gene="HBG1"
          /product="hemoglobin, gamma A"
```

La notación *join* hace referencia a los rangos, dentro de la secuencia, que forman parte del ARNm, esto es, que la cadena de nucleótidos del ARNm será la concatenación de las tres subcadenas indicadas. Por otra parte, la región codificante del gen es:

```
CDS       join(2059..2150,2273..2495,3376..3504)
          /gene="HBG1"
          /codon_start=1
          /product="hemoglobin, gamma A"
          /protein_id="ADL14496.1"
          /db_xref="GI:302313143"
          /translation="MGHFTTEEDKATITSLWGKVNVEDAGGETLGRLL
VVYPWTQRFFDSFGNLSSASAIMGNPKVKAHGKKVLTSLGDATEKHLDDLKGTFAQLSELHCDKLH
VDPENFKLLGNVLVTVLAIHFGKEFTPEVQASWQKMVTAVASALSSRYH"
```

Finalmente, se conocen hasta 44 variaciones del HGB1, cada una con una frecuencia de aparición.

Ahora abordaremos el análisis desde una perspectiva centrada en el gen en lugar de en la secuencia de nucleótidos que lo constituye. La principal ventaja es que la información obtenida está más resumida. Para ello, volvemos a Entrez pero, en lugar de seleccionar la base de datos de nucleótidos, especificamos la base de datos de genes (opción **Gene**) y en la caja de texto de búsqueda escribiremos el identificador del gen HBG1 y hacemos clic en **Go**.

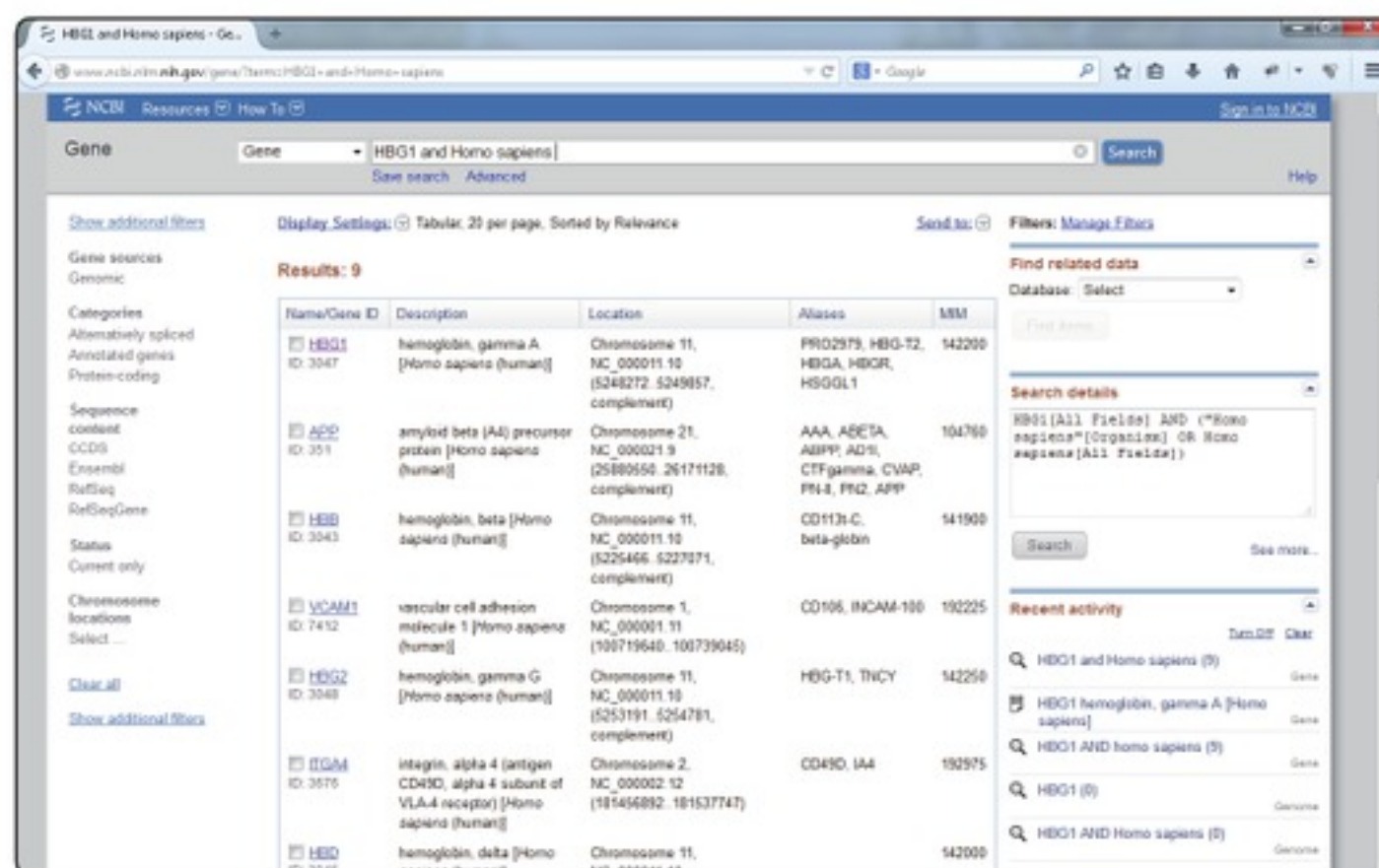


Figura 7.12. Información contenida en la base de datos Gene

En la parte superior de los resultados (ver Figura 7.13) aparece una descripción general del gen, las funciones que realiza y un conjunto de enlaces a otras bases de datos o ficheros del NCBI.

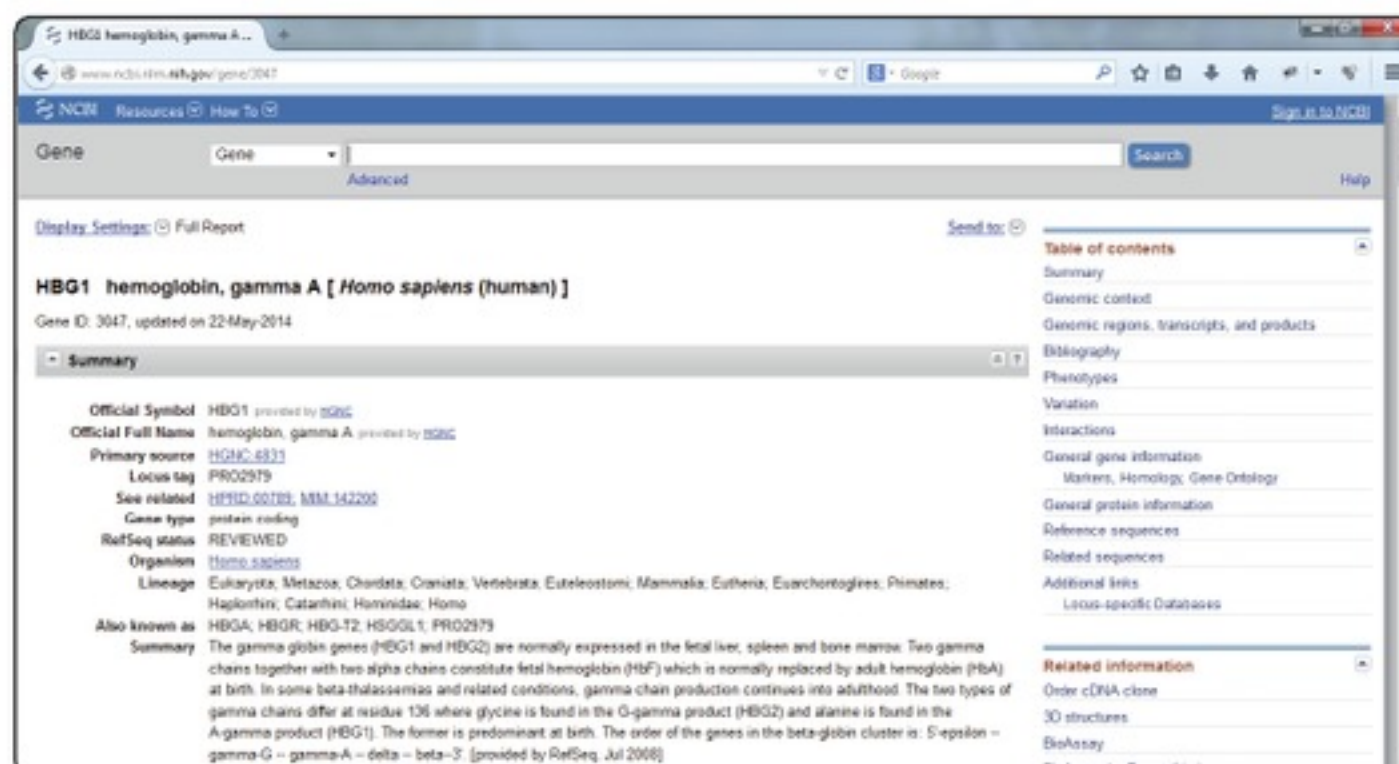


Figura 7.13. Información en Gene sobre el registro buscado

A continuación, las secciones **Genomic context** y **Genomic regions, transcripts and products** muestran una vista esquemática de la estructura del gen y su localización en el cromosoma 11.

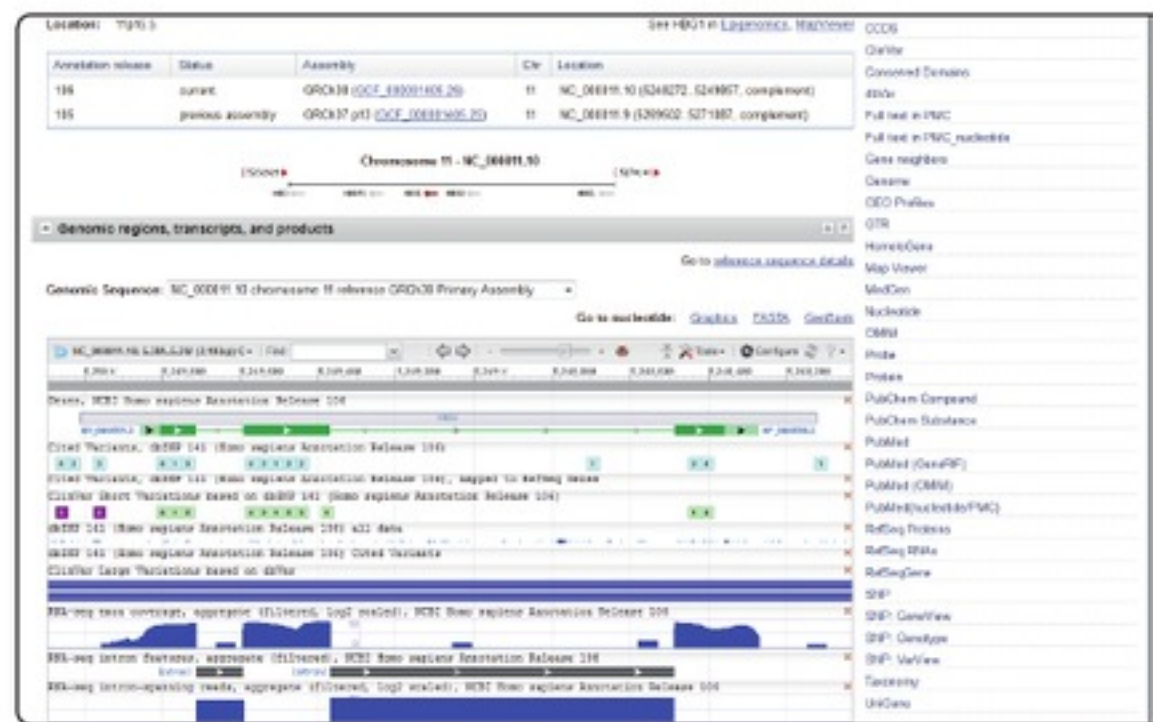


Figura 7.14. Visión genómica del registro

Haciendo clic en el **Map Viewer** se mostrará una ventana con la estructura detallada del gen. Como su nombre indica, MapViewer está diseñado para simplificar la construcción de mapas genómicos.

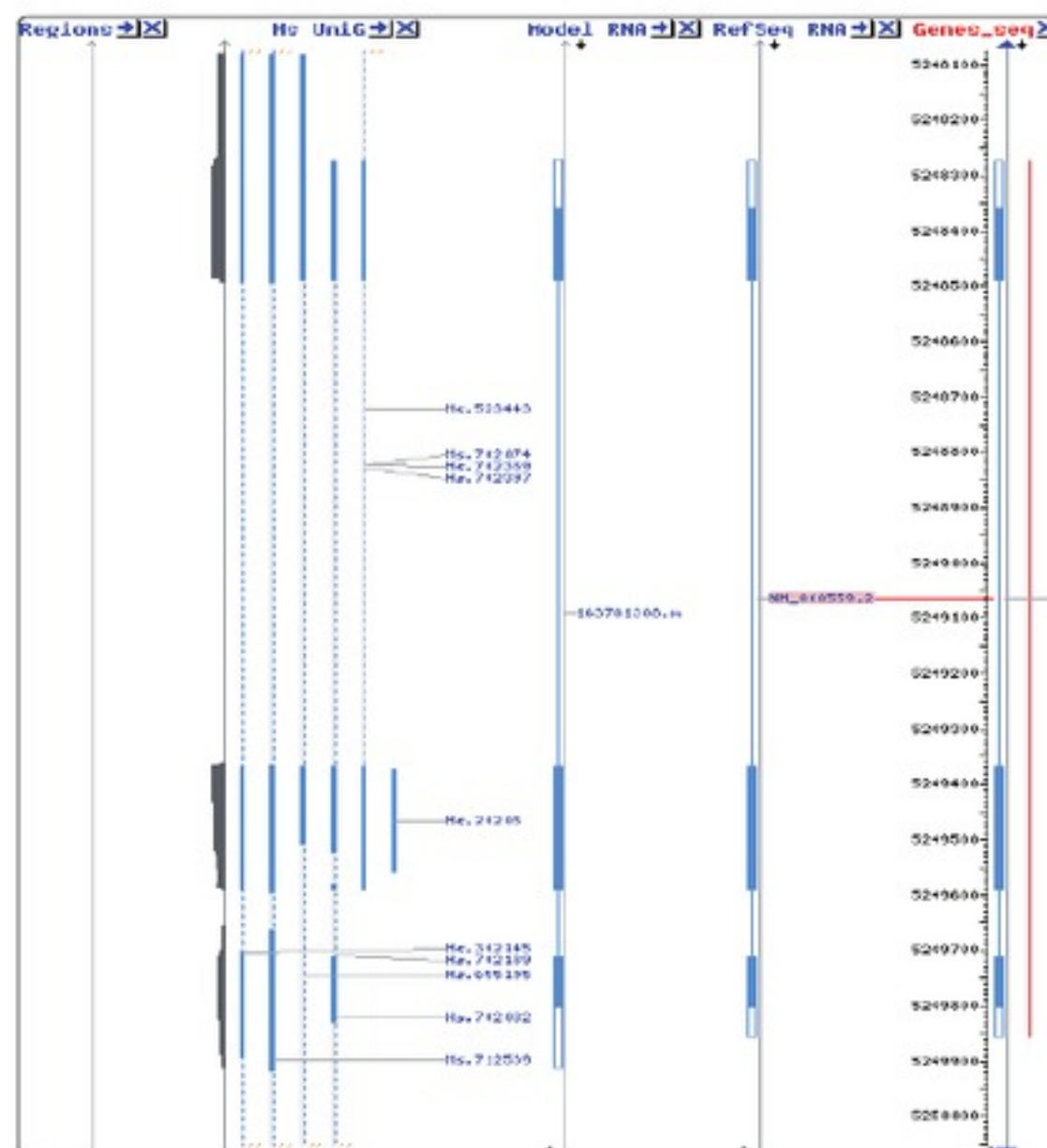


Figura 7.15. MapViewer (I)

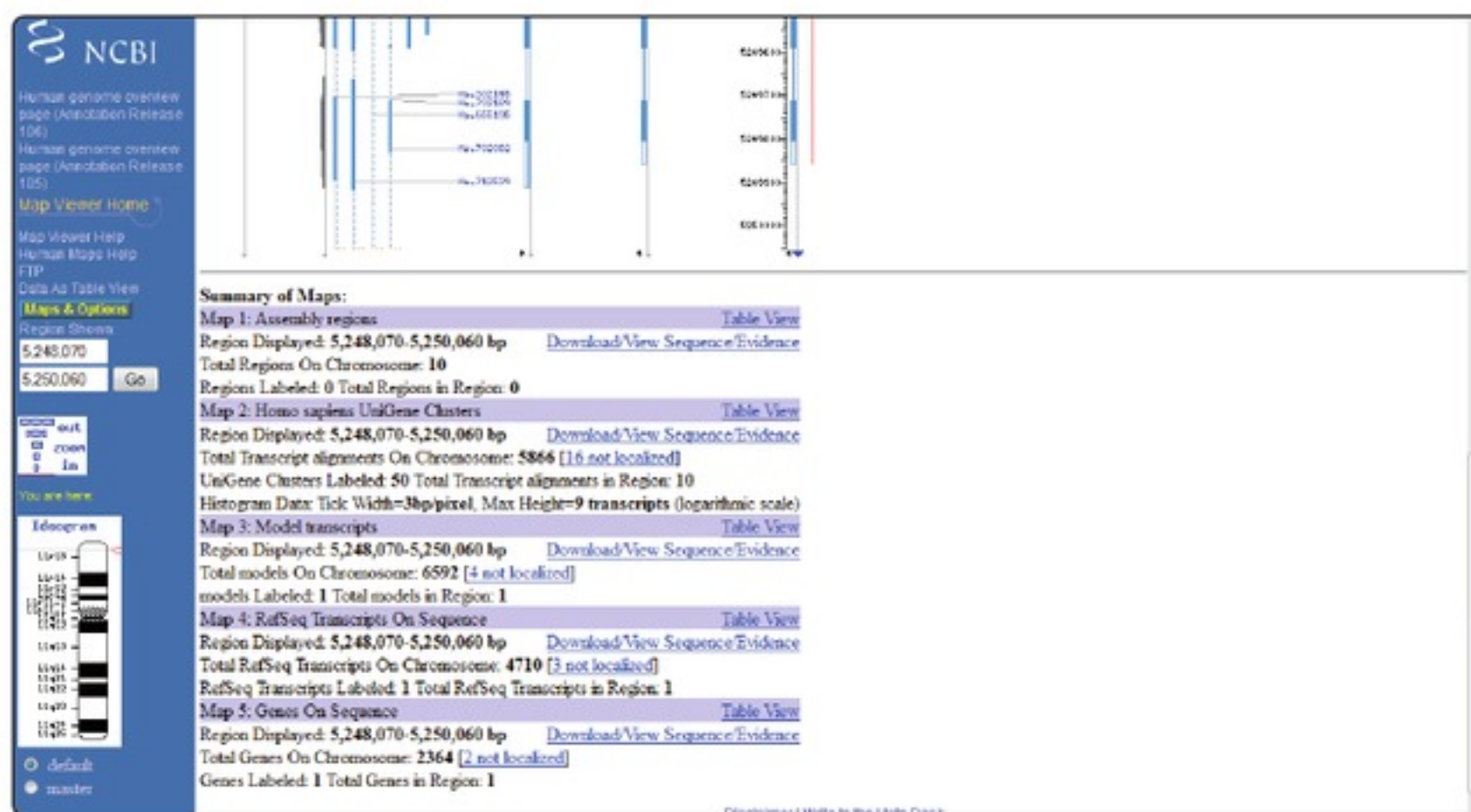


Figura 7.16. Map Viewer (y II)

ANÁLISIS DE SECUENCIAS

El caso más habitual en Biología Molecular consiste en estudiar fragmentos desconocidos de ADN obtenidos como resultado de un experimento de laboratorio. A partir de la secuencia, se trata de identificar qué genes están presentes en la misma. En organismos procariotas la densidad de genes es muy alta, por lo que la tarea es relativamente sencilla. Sin embargo, en organismos eucariotas, en donde las regiones codificantes se encuentran dispersas e imbricadas con exones e intrones, esta tarea es mucho más compleja. Además, existen dificultades añadidas como los genes de ARNm o la densidad de genes de la secuencia anteriormente comentada (por ejemplo, el 80% en bacterias y entre el 3% y el 5% en ser humano). Una vez que se ha identificado el gen, la secuencia de ADN se convierte en secuencia proteica y el problema, entonces, se traslada al modelado de la proteína que tiene por objetivo determinar su estructura y funciones.

En este capítulo centraremos la atención en el análisis de secuencias y dejaremos el modelado de proteínas y la identificación de genes para más adelante. Nuestro punto de partida, por tanto, será realizar el trabajo en el laboratorio y enviar la muestra a la plataforma de secuenciación; el investigador se enfrenta a una tarea de procesamiento de una cantidad enorme de datos. La Figura 8.1 muestra el flujo de trabajo básico de un proyecto de secuenciación genérico. Después de la preparación de las muestras, estas se secuencian en una determinada plataforma. Una vez obtenida la secuencia, se realiza un análisis de calidad y se alinea con una secuencia de referencia con el fin de identificar las variaciones de la secuencia bajo estudio, de tal manera que las mutaciones encontradas se anotarán, se filtrarán y se priorizarán y los resultados se mostrarán en herramientas de visualización. Finalmente, dichos resultados se validarán en el laboratorio.



Figura 8.1. Flujo de trabajo de un proyecto de secuenciación genómico

8.1 DETECCIÓN DE ORF

El primer paso en la identificación de un gen a partir de una secuencia de ADN es determinar la trama de lectura correcta. Puesto que cada aminoácido es un triplete de bases, existen tres posibles tramas de lectura en cada hebra de la molécula de ADN. La trama de lectura correcta u ORF (*Open Reading Frame*) es la trama más larga ininterrumpida por un codón de fin (TGA, TAA o TAG). Cuanto mayor sea la longitud de la ORF, más probabilidad hay de que se trate de un gen.



ORF de entorno a 300 nucleótidos (100 aminoácidos) suelen ser suficientes para la identificación de la mayoría de los genes procariotas.

8.2 ANÁLISIS DE CALIDAD

Una vez que ya se ha obtenido la secuencia correspondiente a la muestra preparada en el laboratorio, el siguiente paso es evaluar la calidad de los datos en bruto y eliminar, recortar o corregir las lecturas que no cumplan con los estándares de calidad adecuados. En efecto, la secuenciación de un fragmento de ADN implica su purificación, clonación utilizando un determinado vector (por ejemplo, un plásmido), amplificación en organismo huésped (generalmente, una bacteria o un virus) y, finalmente, la preparación de la muestra como paso previo a su envío a la plataforma de secuenciación. Durante este proceso, es posible que fragmentos accesorios de ADN se enlacen deliberadamente con la muestra objetivo, todo ello sin olvidar los errores que pudieran ocurrir durante el proceso.

En definitiva, el resultado final es que los datos en bruto generados por la plataforma de secuenciación hay que filtrarlos y recortarlos para prevenir conclusiones biológicas erróneas.

Una de las herramientas más extendidas para determinar y combatir los efectos de la contaminación de secuencias es VecScreen, accesible desde la web del NCBI y que exploraremos en la Práctica 3.

8.3 ALINEAMIENTO

Una vez que se dispone de una secuencia que cumple con los estándares de calidad, el paso siguiente es el alineamiento o comparación de secuencias. La comparación de secuencias es una de las tareas más complicadas, hasta tal punto que incluso hoy en día está pendiente de resolver. Probablemente sea este el motivo que explique la falta de un único método de comparación de secuencias. En general, los algoritmos de alineamiento analizan secuencias de caracteres considerando, especialmente, algunas características de las secuencias biológicas:

- a) Las secuencias biológicas tienen un juego de caracteres definidos (códigos IUPAC-IUB).
- b) Existen caracteres que representan el valor de dos nucleótidos simultáneamente.

La Tabla 8.1 resume los métodos más utilizados y que serán estudiados más adelante en este capítulo.

Herramienta	Situaciones en que se utiliza
Gráfico de puntos	• Exploración general de la secuencia • Búsqueda de repeticiones • Búsqueda de variaciones largas • Extracción de fragmentos para alineamientos múltiples
Alineamiento local	• Comparación de secuencias parcialmente homólogas • Alineamientos de alta calidad • Análisis residuo por residuo
Alineamiento global	• Comparación de dos secuencias en toda su longitud • Identificación de variaciones • Comprobación de la calidad de los datos

Tabla 8.1. Herramientas de comparación de secuencias

El alineamiento de secuencias es una manera de comparar dos secuencias primarias de ADN, ARN o proteína para identificar regiones similares que tengan su justificación en una relación funcional, estructural o evolutiva.

En general, existen dos escenarios diferentes: el alineamiento de la secuencia con una de referencia o bien el alineamiento de una secuencia para la que no existe referencia.

En el primer caso, la secuenciación para una posterior comparación con un genoma de referencia, se siguen tres pasos. Primeramente, las moléculas de ADN se rompen, por posiciones aleatorias, en fragmentos más pequeños y se secuencia cada uno de estos fragmentos utilizando enzimas de restricción o métodos mecánicos. A continuación, se crea una librería que contiene esos fragmentos de tamaño conocido y, finalmente, los fragmentos se mapean siguiendo las instrucciones de una secuencia de referencia conocida. Este proceso queda resumido en la Figura 8.2.

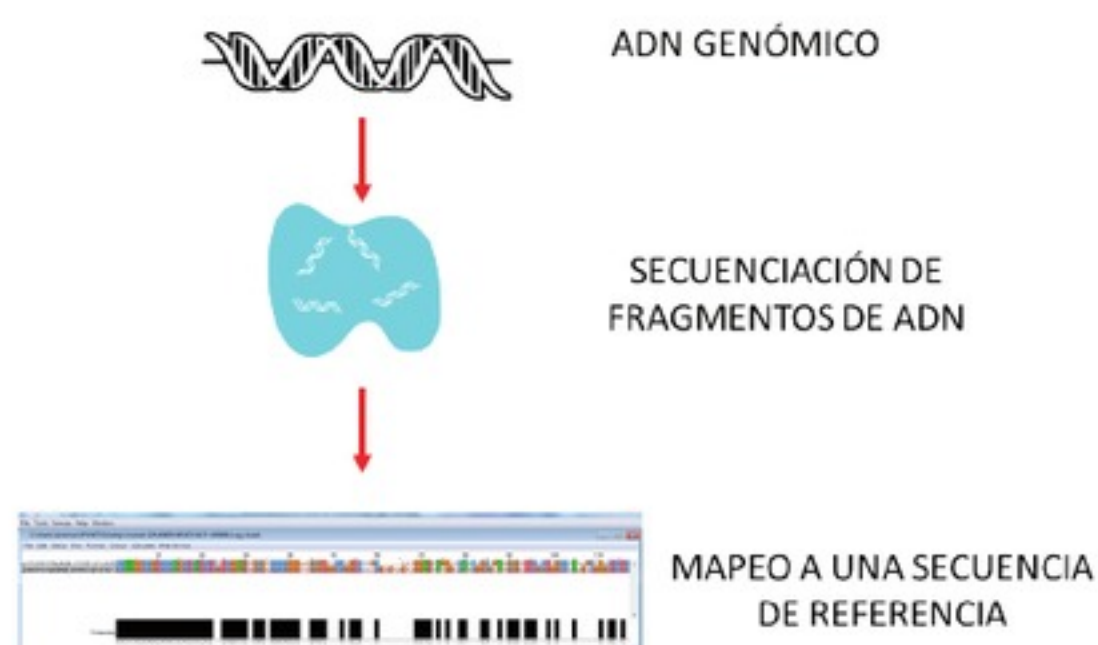


Figura 8.2. Alineamiento con una secuencia de referencia

En el caso del genoma humano, por ejemplo, existen en la actualidad dos grandes librerías de referencia: el proyecto ENCODE de la UCSC (University of Santa Cruz) y la base de datos del GRC (Genome Reference Consortium). Ambos recursos proporcionan varias versiones del genoma humano (hg18 y h19 y GRC36 y GRC37, respectivamente). Realmente, ambas versiones son idénticas y la diferencia estriba en la nomenclatura empleada.

Por otra parte, cuando se trata de una nueva secuencia para la que no existe referencia, los fragmentos de secuencia se ensamblan en grupos solapados llamados *contigs* y la secuencia final (consenso) se obtiene a partir de estos *contigs* sin necesidad de disponer de información sobre la secuencia original (ver Figura 8.3). La precisión de este método aumenta cuando se comparan secuencias largas de moléculas de ADN de baja tasa de repetición.

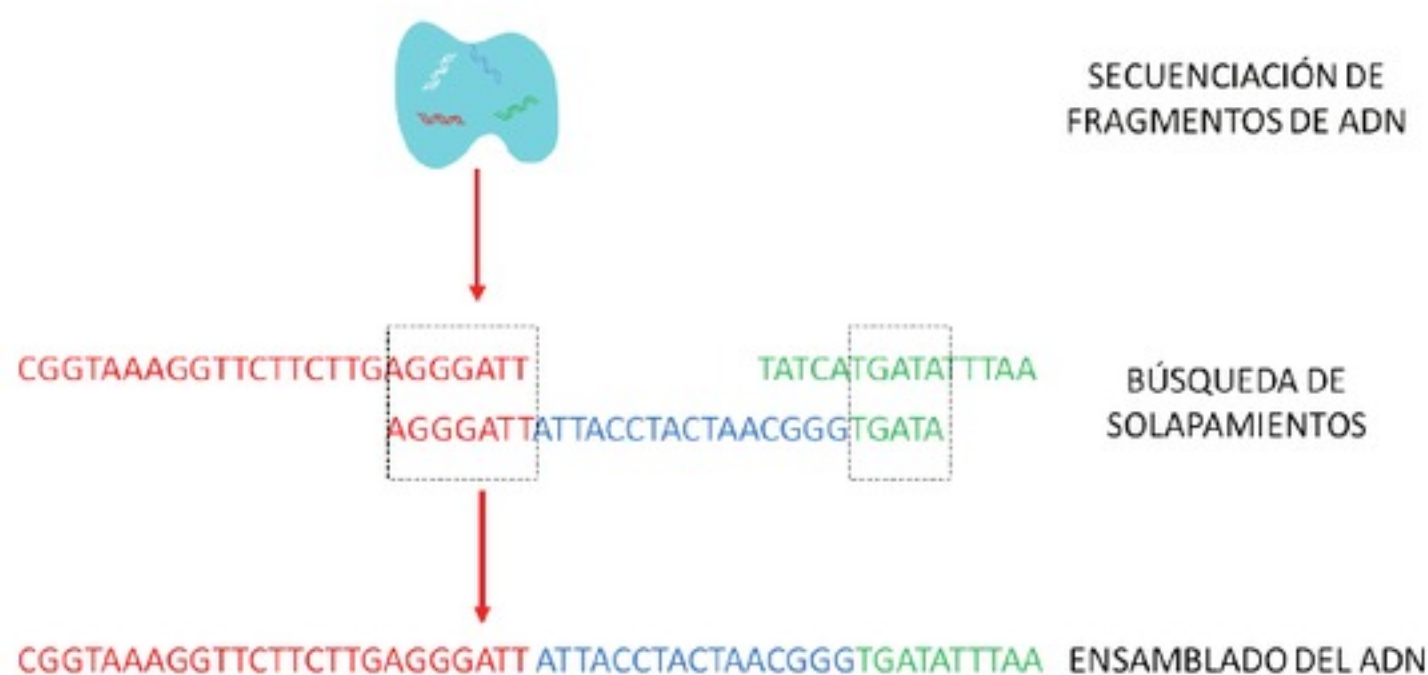


Figura 8.3. Ensamblado de una secuencia de ADN

En cualquiera de los casos, el resultado de la comparación de secuencias es un índice de similitud. Dependiendo del número de secuencias que se comparen, encontramos alineamiento de pares (cuando se comparan dos secuencias) o alineamiento múltiple (más de dos secuencias).

8.3.1 Gráficos de puntos

Una de las maneras más sencillas de comparar dos secuencias es con un método visual llamado gráfico de puntos. Se dispone una secuencia en sentido horizontal y la otra en sentido vertical (ver Figura 8.4) formando una matriz. Cuando exista una coincidencia de nucleótidos, se marca la casilla correspondiente de la matriz.



Cuando se vayan a comparar dos secuencias, es recomendable empezar realizando un gráfico de puntos comparando cada una de ellas consigo misma ya que de esta manera se simplifica la identificación de regiones repetidas dentro de la propia secuencia.

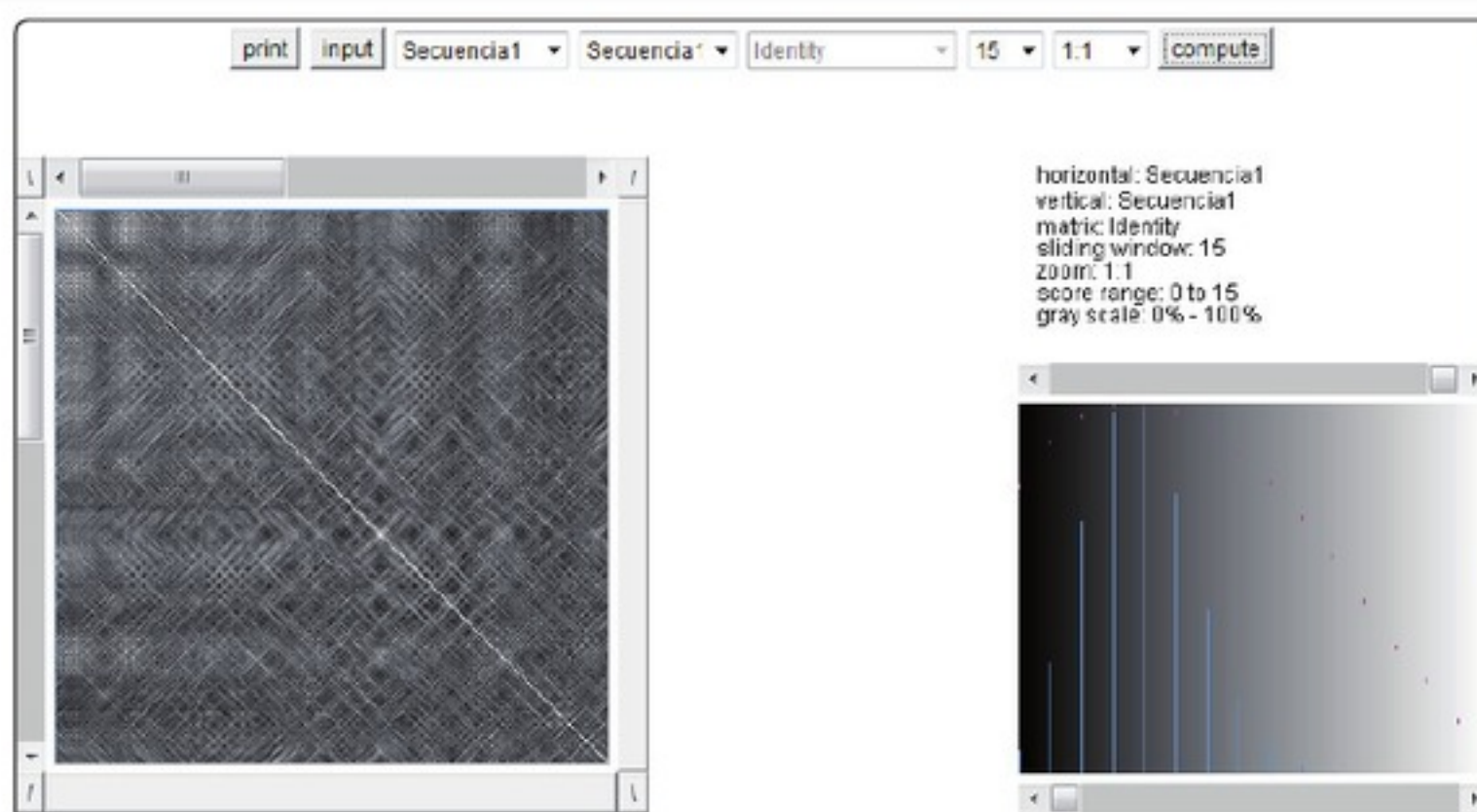


Figura 8.4. Ejemplo de gráfico de puntos

Aunque el gráfico de puntos es sencillo, para secuencias largas se vuelve inmanejable desde el punto de vista práctico. Para solventar estos inconvenientes se emplean los alineamientos. Básicamente, existen dos tipos de alineamiento. En el alineamiento global, ambas secuencias se alinean en toda su longitud y se encuentra así el mejor alineamiento posible. Sin embargo, en el alineamiento local, se busca el mejor alineamiento entre las regiones más similares y se ignora el resto.

8.3.2 Alineamiento de pares

El alineamiento de pares de secuencias (PSA, *Pairwise Sequence Alignment*) consiste en comparar dos secuencias biológicas (ácidos nucleicos o proteínas) con el fin de obtener la similitud u homología de ambas, determinando en qué se parecen y en qué se diferencian.

La herramienta de alineamiento de secuencias más utilizada es BLAST (*Basic Local Alignment Search Tool*). BLAST comprueba si la secuencia de entrada tiene alguna similitud con una librería de secuencias, comparando dicha entrada

con las secuencias de una base de datos y calculando el grado de relevancia de las coincidencias entre las secuencias. En la Práctica 2 estudiaremos este programa con más detalle.

Probablemente, la versión más popular de BLAST es la implementación del NCBI, accesible desde <http://www.ncbi.nlm.nih.gov/BLAST/>. Sin embargo, existen otras muchas que pueden encontrarse fácilmente lanzando una búsqueda sencilla en cualquier buscador.

Existen varias versiones de BLAST en el NCBI, cada una de las cuales orientada a un fin específico:

- BLASTN: se emplea para comparar una secuencia de nucleótidos bajo estudio con la información guardada en una base de datos de secuencias de nucleótidos.
- BLASTP: compara una secuencia de aminoácidos con una base de datos de proteínas, es decir, resulta útil cuando, teniendo una secuencia proteica, se desea encontrar secuencias proteicas similares en una base de datos de secuencias.
- BLASTX: traduce una secuencia de ADN bajo estudio en sus 6 proteínas diferentes (cada una con una ORF) y compara cada una de estas proteínas con las guardadas en una base de datos de secuencias de proteínas.
- TBLASTN: compara la secuencia de aminoácidos bajo estudio con una base de datos de nucleótidos traducida en sus 6 posibles ORF.
- TBLASTX: compara las seis traducciones en sus marcos de lectura de la secuencia de nucleótidos bajo estudio, contra las seis traducciones en sus marcos de lectura de toda la base de datos de nucleótidos.

8.3.3 Alineamiento múltiple

El alineamiento de pares es útil a la hora de buscar secuencias similares en una base de datos. Sin embargo, cuando se requiere comparar un número mayor de secuencias simultáneamente, el PSA se vuelve inabordable. En estos casos, se emplea el alineamiento múltiple (MSA, *Multiple Sequence Alignment*) que permite que, de una sola tacada, se puedan comparar varias secuencias. Por otra parte, el alineamiento múltiple, en lugar de consultar una base de datos y generar un *ranking* de secuencias por orden de similitud, compara y alinea las secuencias de entrada. El alineamiento múltiple consiste en reescribir las secuencias de entrada de manera que los fragmentos similares se encuentren en la misma columna. De esta manera, es posible detectar características comunes en conjuntos de secuencias y predecir la estructura y la función de las mismas.



Incluso aunque se esté interesado en la similitud entre dos secuencias, resulta conveniente realizar un alineamiento múltiple con todas las secuencias disponibles (o un número razonable de ellas). La utilización de estas secuencias adicionales mejora la exactitud de los alineamientos entre pares de secuencias.

En general, el proceso de alineamiento múltiple se basa en la repetición sistemática de los siguientes pasos:

- Obtención de las secuencias a través de una búsqueda en base de datos u otro medio.
- Localización de las regiones que interesa alinear.
- Ejecutar el programa de alineamiento. Existe gran cantidad de software de alineamiento múltiple (TCoffee, MUSCLE, MAFFT, PROBCONS, etc.). Sin embargo, el más ampliamente utilizado, y que será desarrollado en la Práctica 3, es Clustal Omega.
- Revisar manualmente los resultados del alineamiento, prestando especial atención a las regiones con huecos.
- Eliminar las secuencias que parezcan distorsionar los resultados y volver a alinear.
- Una vez identificados los residuos clave en las secuencias que han pasado el filtro del punto anterior, añadir el resto de secuencias intentando preservar las características clave de la familia.

8.3.4 Puntuación del alineamiento

Para poder comparar alineamientos es necesario establecer algún sistema de puntuación que cuantifique la similitud entre la secuencia bajo estudio y una secuencia de referencia. De esta manera, el mejor alineamiento será el que presente una mayor puntuación.

En el caso de secuencias de nucleótidos, el sistema de puntuación resulta relativamente sencillo puesto que únicamente hay que tener en cuenta tres situaciones simples: coincidencia (*match*), no coincidencia (*mismatch*) y hueco

(*gap*). Supongamos que queremos obtener la puntuación del alineamiento entre las dos siguientes secuencias:

```
gg-a-tcgga--tc  
ggaaatcggaaatc
```

Y que el sistema de puntuación elegido es tal que la coincidencia puntúa 1, la no coincidencia puntúa 0 y el hueco puntúa -1. La puntuación del alineamiento sería, por tanto:

```
Puntuación = 10 coincidencias * 1 + 0 no coincidencias * 0 + 4  
huecos * (-1) = 6
```

Existen otros sistemas de puntuación de alineamiento más complejos que incluyen penalizaciones, no solo para los huecos, sino también para las situaciones de apertura o extensión de los mismos.

En el caso de cadenas de aminoácidos la situación se vuelve más compleja puesto que la comparación no es binaria sino que dos aminoácidos pueden ser iguales, distintos o tener cierto parecido. La solución consiste en utilizar matrices de sustitución. Las matrices de sustitución contienen valores proporcionales a la probabilidad de que un aminoácido X sea sustituido por un aminoácido Z, y se calcula la matriz para todos los pares de aminoácidos posibles.

No existe una matriz de sustitución única que pueda emplearse siempre sino que se utiliza una u otra en función de la familia de proteínas y del grado de similitud u homología esperado. Sin embargo, las más extendidas son las matrices PAM y las matrices BLOSUM.

Las matrices PAM (*Percent Accepted Mutation Matrix*) se derivan de alineamientos globales de secuencias. Una matriz PAM es una matriz cuadrada de orden 20, puesto que 20 son los aminoácidos esenciales. El valor de cada elemento de la matriz está relacionado con la probabilidad de que el aminoácido correspondiente a la columna antes de la mutación se alinee con el aminoácido correspondiente a la fila. De esta manera, es posible caracterizar la distancia evolutiva entre las secuencias, de manera que cuanto mayor sea este número, mayor será dicha distancia. Es decir, la matriz PAM250 se utilizará para secuencias de proteínas cuyo grado de homología esperado sea bajo, mientras que emplearemos la matriz PAM60 cuando se espera un parecido de en torno al 60% entre las secuencias.

Por su parte, las matrices BLOSUM se construyen a partir de alineamientos locales. Al igual que las matrices PAM, las matrices BLOSUM se caracterizan por

un número. Si el número es alto es porque la matriz se ha diseñado para comparar secuencias cercanas evolutivamente mientras que, por el contrario, si el número es bajo, la matriz está orientada a secuencias lejanas.

Como vemos, ambos tipos de matrices tienen el mismo objetivo pero utilizan metodologías distintas. Mientras que PAM deduce información evolutiva a partir de la proximidad entre secuencias, BLOSUM se centra en analizar las mutaciones entre secuencias relacionadas. No obstante, es posible establecer cierta equivalencia, como muestra la Tabla 8.2:

PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM60
PAM200	BLOSUM52
PAM250	BLOSUM45

Tabla 8.2. Equivalencia entre las matrices PAM y las matrices BLOSUM



El cálculo de las matrices de sustitución PAM y BLOSUM para la comparación de secuencias queda fuera del ámbito de este libro. Remitimos a la Bibliografía, en donde se encontrarán referencias que detallan el proceso con gran exactitud.

8.4 IDENTIFICACIÓN DE VARIACIONES

Una parte crucial del análisis de secuencias es la identificación de variaciones. Una variación genética es una diferencia entre la secuencia bajo estudio y la secuencia de referencia. Pueden deberse o bien a la herencia de uno de los progenitores o bien a una mutación que introduce un cambio en el cromosoma heredado. Estas variaciones afectan a la fabricación de proteínas y pueden tener efecto sobre el fenotipo o no.

La forma de variación más sencilla es aquella en la que las secuencias difieren en un único nucleótido. Este tipo de variación estructural recibe el nombre de SNP (*Single Nucleotide Polymorphism*) o SNV (*Single Nucleotide Variation*). Los SNP son sustituciones de un único nucleótido y se dividen en dos tipos: transiciones

(intercambio de dos purinas o dos pirimidinas) y transversiones (intercambios entre purinas o pirimidinas). La Figura 8.5 muestra un ejemplo de transición:

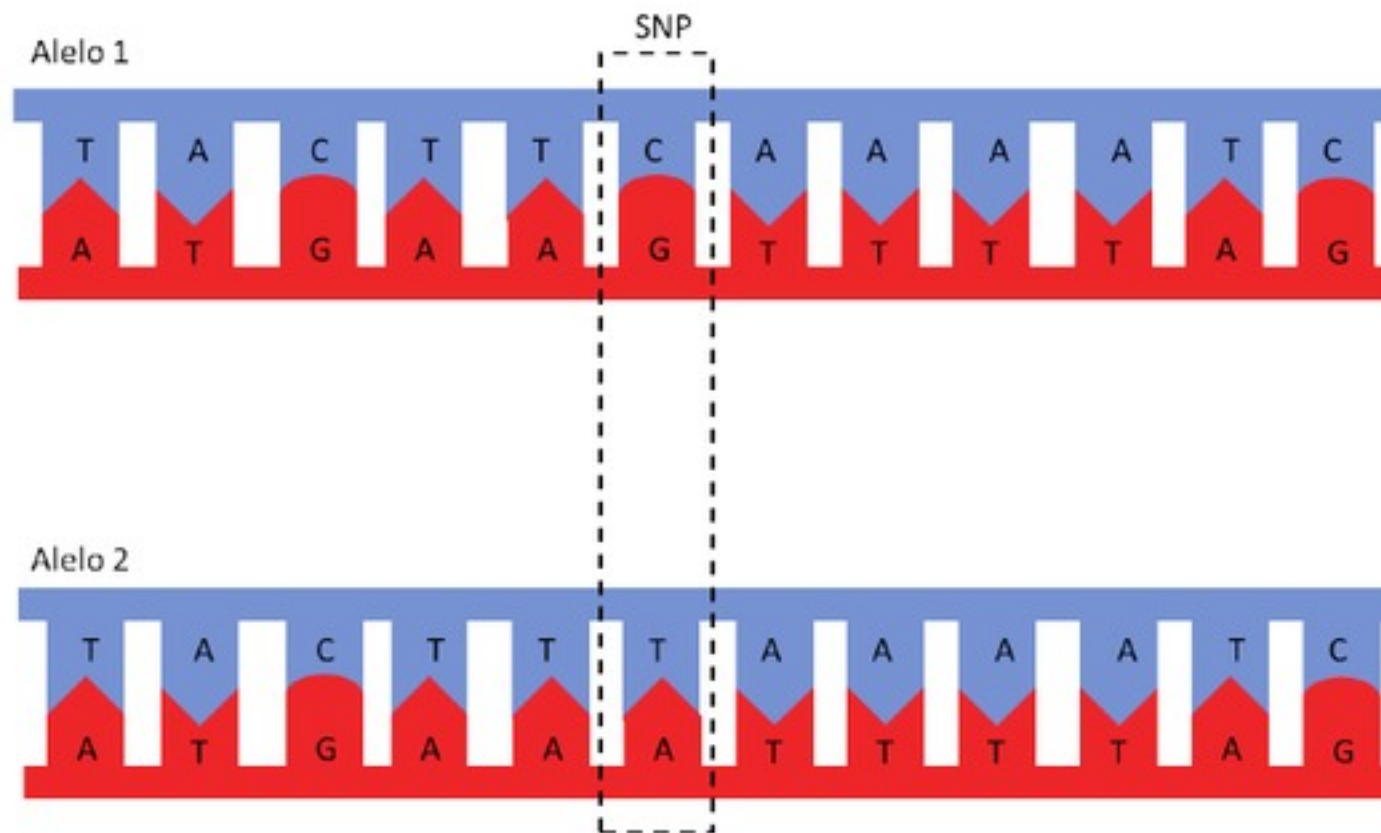


Figura 8.5. Ejemplo de SNP

Existen bases de datos públicas sobre SNP. Dos de las más utilizadas son el dbSNP del NCBI, que ofrece información sobre variaciones de distintas especies; y, el HGMD (*Human Gene Mutation Database*), especializada en mutaciones del genoma humano asociadas a enfermedades y SNP funcionales.

Sin embargo, las variaciones estructurales no tienen porqué ser de un único nucleótido. De hecho, en la actualidad existe una tendencia al estudio de variaciones polinucleótidas, también llamadas variomas estructurales y que pueden ser:

- Sustituciones de una base por otra: también se llaman mutaciones puntuales o *indels* (*insertion and deletion*). Generalmente, son de un único nucleótido.
- Inserciones: se introducen bases adicionales en la secuencia.
- Borrados: se eliminan una o más bases de la secuencia original.
- Inversiones: parte de la secuencia aparece invertida cambiando la estructura del cromosoma.
- Translocaciones: ocurre cuando un segmento del cromosoma se desplaza a un nuevo lugar del genoma.

Además, las variaciones pueden clasificarse atendiendo a su ámbito en cromosómicas, cuando se producen a nivel de un cromosoma; y, génicas, si tienen lugar a nivel de gen. Si la variación ocurre dentro de un gen un posible efecto es una variación en la pauta de lectura durante la fase de traducción, sobre todo si la mutación afecta a una o dos bases, ya que esto tiene un impacto directo en la detección del codón de inicio o de fin. Si, por el contrario, la mutación afectara a tres bases, el codón quedaría intacto y el resultado sería la adición o eliminación de un aminoácido.

Otro tipo de variaciones son los polimorfismos. Se trata de variaciones en la secuencia entre los individuos de una población, algunos de ellos no tienen manifestaciones fenotípicas pero otros sí, por ejemplo, como el gen que codifica el color del pelo o de los ojos. Es lo que se conoce como CNP (*Copy Number Polymorphism*). Si el polimorfismo es de un solo nucleótido, recibe el nombre SNP.

Por otra parte, los variomas estructurales se dividen en dos grandes categorías (ver Figura 8.6): balanceados, cuando hacen referencia a reordenaciones que no modifican la longitud total de la secuencia (inversiones o translocaciones intra o intercromosómicas); y, no balanceados, cuando los variomas sí que modifican la longitud total de la secuencia (inserciones y delecciones). Los variomas estructurales no balanceados también se llaman CNV (*Copy Number Variations*).

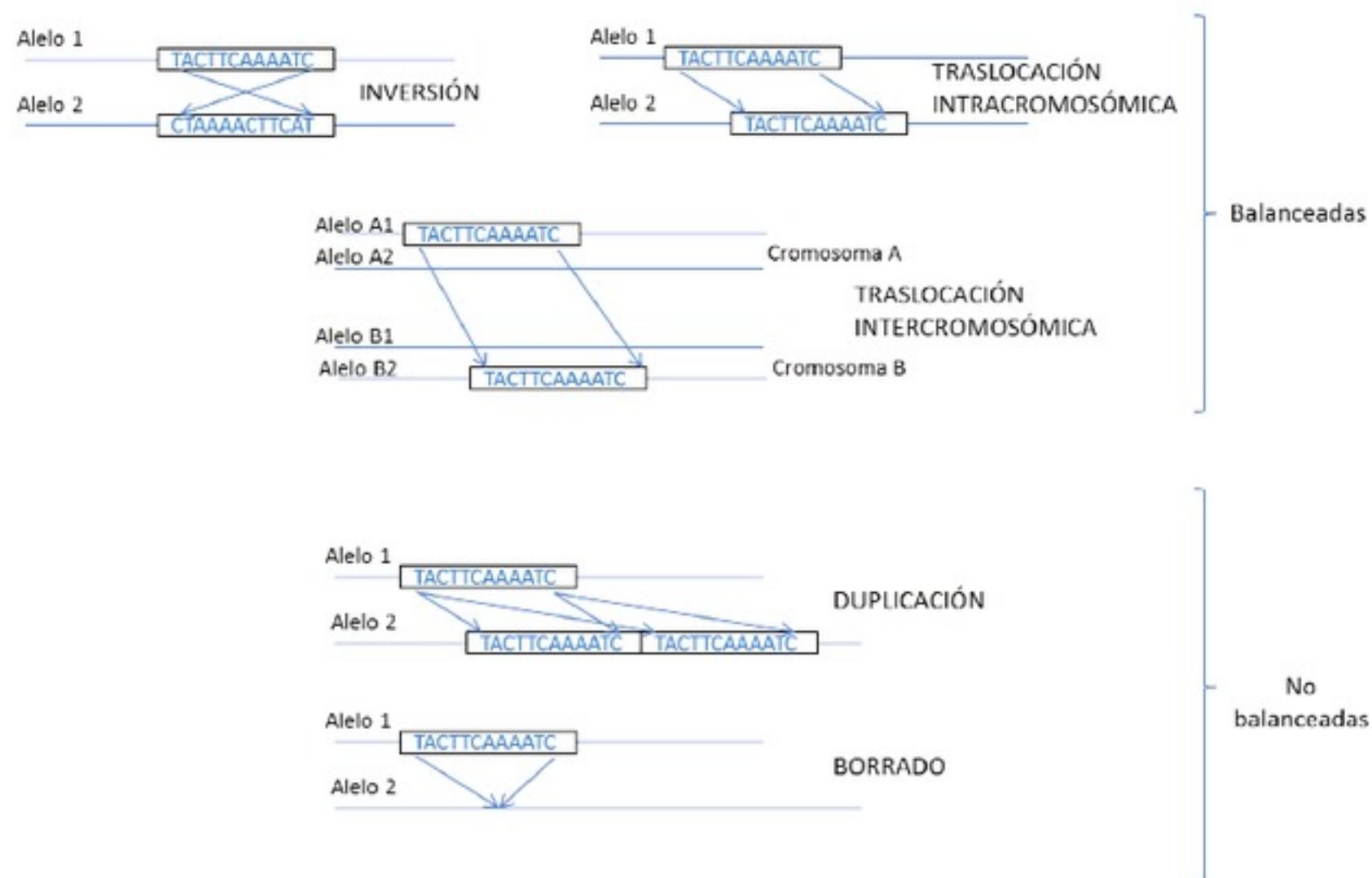


Figura 8.6. Variaciones estructurales

Uno de los problemas de las bases de datos genómicas actuales es, precisamente, la caracterización de variaciones. A grandes rasgos, una variación queda descrita por la posición en que se ha localizado, el valor original que tenía la secuencia de referencia y el valor de la variación. Desgraciadamente, estos parámetros dependen del algoritmo de alineamiento utilizado y, como consecuencia, científicos diferentes podrían llegar a conclusiones distintas sobre el mismo hecho. La Figura 8.7 ilustra este hecho:



Figura 8.7. Problemas de los algoritmos de alineamiento

En el caso del primer algoritmo de alineamiento, se detecta una sustitución de dos bases, mientras que según el segundo, existe una inserción en una posición. Sin embargo, se trata del mismo cambio en la secuencia.

Una posible solución a este tipo de problemas es la caracterización de las variaciones mediante regiones adyacentes (*flanking sequences*), en lugar de basarse únicamente en la posición del nucleótido. De esta manera, una región adyacente estará formada por un número suficiente de bases que permiten la identificación de una secuencia dada. Con esta nueva aproximación (ver Figura 8.8), no existe ambigüedad posible y se detecta la sustitución en cualquier caso.

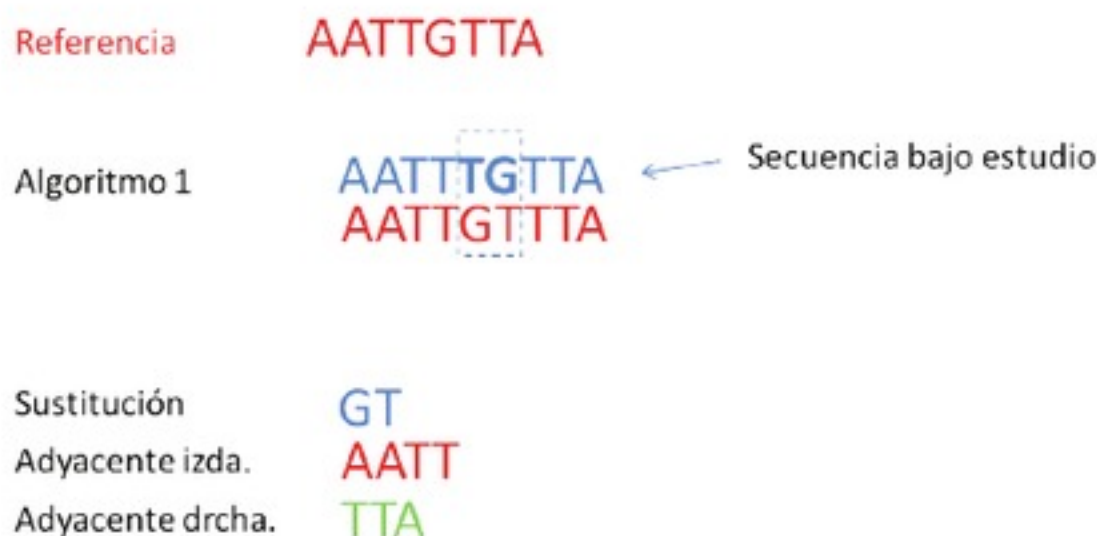


Figura 8.8. Alineamiento mediante regiones adyacentes

Sin embargo, la utilización de regiones adyacentes plantea un problema y es determinar el tamaño óptimo tal que se garantice la identificación de la secuencia.

8.5 ANOTACIÓN

El término anotación hace referencia a la obtención de información biológica de datos de secuencias sin procesar. Una anotación es estructural cuando está relacionada con la identificación de genes y otros elementos; mientras que una anotación es funcional, si está orientada a averiguar su función en el organismo. Las anotaciones se consiguen gracias a la aplicación de las tecnologías de la información a la gestión y análisis de datos biológicos.

La mayoría de herramientas de anotación de variaciones se centran en la anotación de SNP, puesto que es este tipo de variaciones el más fácil de identificar y analizar. No obstante, algunas herramientas de última generación son capaces de anotar CNV. La anotación más común es proporcionar enlaces a bases de datos de variaciones como dbSNP. El número de variaciones depende del genoma. Por ejemplo, para el genoma humano actualmente existen entre 3 y 5 millones de variaciones.

Hay que subrayar que la mayoría de estudios actuales se centran en el exoma y no en el genoma completo, en parte, porque la secuenciación del exoma tiene un coste asociado menor. Por otro lado, se cree que las variaciones del exoma pueden tener un mayor impacto funcional en las enfermedades humanas. Sin embargo, estudios recientes han demostrado que también las regiones no codificantes influyen en las enfermedades.

8.6 VISUALIZACIÓN

Las herramientas de visualización de datos genómicos se clasifican en:

- Herramientas con soporte para tratamiento de secuencias nuevas o experimentos de secuenciación.
- Navegadores genómicos: muestran los datos de secuenciación juntos con las anotaciones que se hayan realizado sobre los mismos en una interfaz gráfica unificada.

- Visores comparativos que facilitan la comparación de secuencias de múltiples organismos o individuos.

8.7 PIPELINES ANALÍTICOS Y SISTEMAS DE FLUJO DE TRABAJO

La combinación de estos métodos de análisis para la obtención de resultados con significado biológico es todavía un reto, incluso para usuarios avanzados.

Una alternativa viable es la utilización de *pipelines* capaces de analizar todos los pasos y que abarcan desde el procesamiento de las secuencias en bruto hasta la anotación de variaciones. Generalmente, establecen un orden predefinido de las etapas de análisis e incorporan algoritmos que se pueden modificar o reemplazar para adaptarlos a unas necesidades concretas.



PRÁCTICA 3: ANÁLISIS DE SECUENCIAS

9.1 ANÁLISIS DE LA CALIDAD CON VECSCREEN

VecScreen es un programa del NCBI que sirve para determinar si una muestra está contaminada utilizando un algoritmo de similitud de secuencias con las almacenadas en la base de datos UniVec.



Cuando se recibe una secuencia procedente de una plataforma de secuenciación hay que recordar que existen dos potenciales fuentes de error: las inherentes a las limitaciones tecnológicas de las máquinas secuenciadoras y las debidas a la contaminación de la secuencia en el laboratorio, previamente a su envío al secuenciador.

En primer lugar, abrimos un navegador web y tecleamos la dirección de la página principal de VecScreen: <http://www.ncbi.nlm.nih.gov/tools/vecscreen/>. Seguidamente, copiamos y pegamos las siguientes secuencias en el área de texto de la pantalla.

```
>Secuencial Libro Bioinformatica
GTTGCCCCGGCCGCCGCCGCCGCCGCGCTACCTATCTCGCCGCCGCCGCCCTCGACC CCTGCTT
CCGTCTGGCTGCCTGTGCCGCGTGGTGCCGGACCCGGGGCAGTGTGCAGGGCCGCCGGGAAAGG
GAAGGAGGTGCTCAGCGGCGTGGTCTTCCAGCCATTCGAGGAGCTCAAGGGGGAGCTCTCCCTC
GTCCCCCAGGCCAAGGACCAGTCTCTCGCTAGGCAAAAGTTCGTGACGAGTGCGAGGCCGCCA
TCAACGAGCAGATCAATGTGGAGTACAATGCATCGTACGCGTACCACTCCCTTTTCGCCTACTT
TGATCGTGACAACGTTGCTCTCAAGGGATTCGCCAAATTCTTCAAAGAATCCAGCGATGAGGAG
AGGGATCACGCAGAGAACTCATCAAGTACCAGAACATGCGTGGAGGCAGGGTGCGGCTCCAGT
```

```

CCATCGTCACACCTTTGACAGAGTTCGACCATCCTGAGAAAGGGGATGCCTTGTATGCTATGGA
TTGGCCTTGGCTCTCGAAAAGCTTGTAATGAGAAGTTGCACAACCTGCACAGTGTGGCATCAA
GGTGCAATGATCCACAGCTGACCGACTTCGTTGAGAGCGAATTCCTTGAGGAGCAGGTTGAAGC
CATCAAGAAGATCTCTGAGTATGTCGCCCAGCTGAGAAGAGTGGGAAAGGGGCATGGGGTGTGG
CACTTTGATCAGAAGCTGCTTGAGGAAGAAGCT
> Secuencia2 Libro Bioinformatica
CCGTACCTTGNNGCGATTGGGCCCTCTAGATGCATGCTCGAGCGGCCCGCCAGTGTGATGGATA
TCTGCAGAATTCGCCCTTTGCTGCAGCCTTTCCGCCATGCTTCCTCCTAGGGTTGCCCCGGCC
GCCGCCGCCGCCGCCGCTACCTATCTCGCCGCCGCCGCTCGACCCCTGCTTCCGTCTGGCTG
CCTGTGCCGCGTGGTGCCGGACCCGGGGCAGTGTGCAGGGCCGCCGGGAAAGGGAAGGAGGTG
CTCAGCGGCGTGGTCTTCCAGCCATTCGAGGAGCTCAAGGGGGAGCTCTCCCTCGTCCCCCAG
GCCAAGGACCAGTCTCTCGCTAGGCAAAAGTTCGTGACGAGTGCAGAGGCCGCCATCAACGAG
CAGATCAATGTGGAGTACAATGCATCGTACGCGTACCACTCCCTTTTCGCCTACTTTGATCGT
GACAACGTTGCTCTCAAGGGATTGCGCAAATTCTTCAAAGAATCCAGCGATGAGGAGAGGGAT
CACGCAGAGAACTCATCAAGTACCAGAACATGCGTGGAGGCAGGGTGCGGCTCCAGTCCATC
GTCACACCTTTGACAGAGTTCGACCATCCTGAGAAAGGGGATGCCTTGTATGCTATGGAGTTG
GCCTTGGCTCTCGAAAAGCTTGTAATGAGAAGTTGCACAACCTGCACAGTGTGGCATCAAGG
TGCAATGATCCACAGCTGACCGACTTCGTTGAGAGCGAATTCCTTGAGGAGCAGGTTGAAGCC
ATCAAGAAGATCTCTGAGTATGTCGCCCAGCTGAGAAGAGTGGGAAAGGGGCATGGGGTGTGGC
ACTTTGATCAGAAGCTGCTTGAGGAAGAAGCT

```

Figura 9.1. Secuencias bajo estudio

Los formatos permitidos son FASTA o bien un *accession* o un GI (ver Figura 9.2).

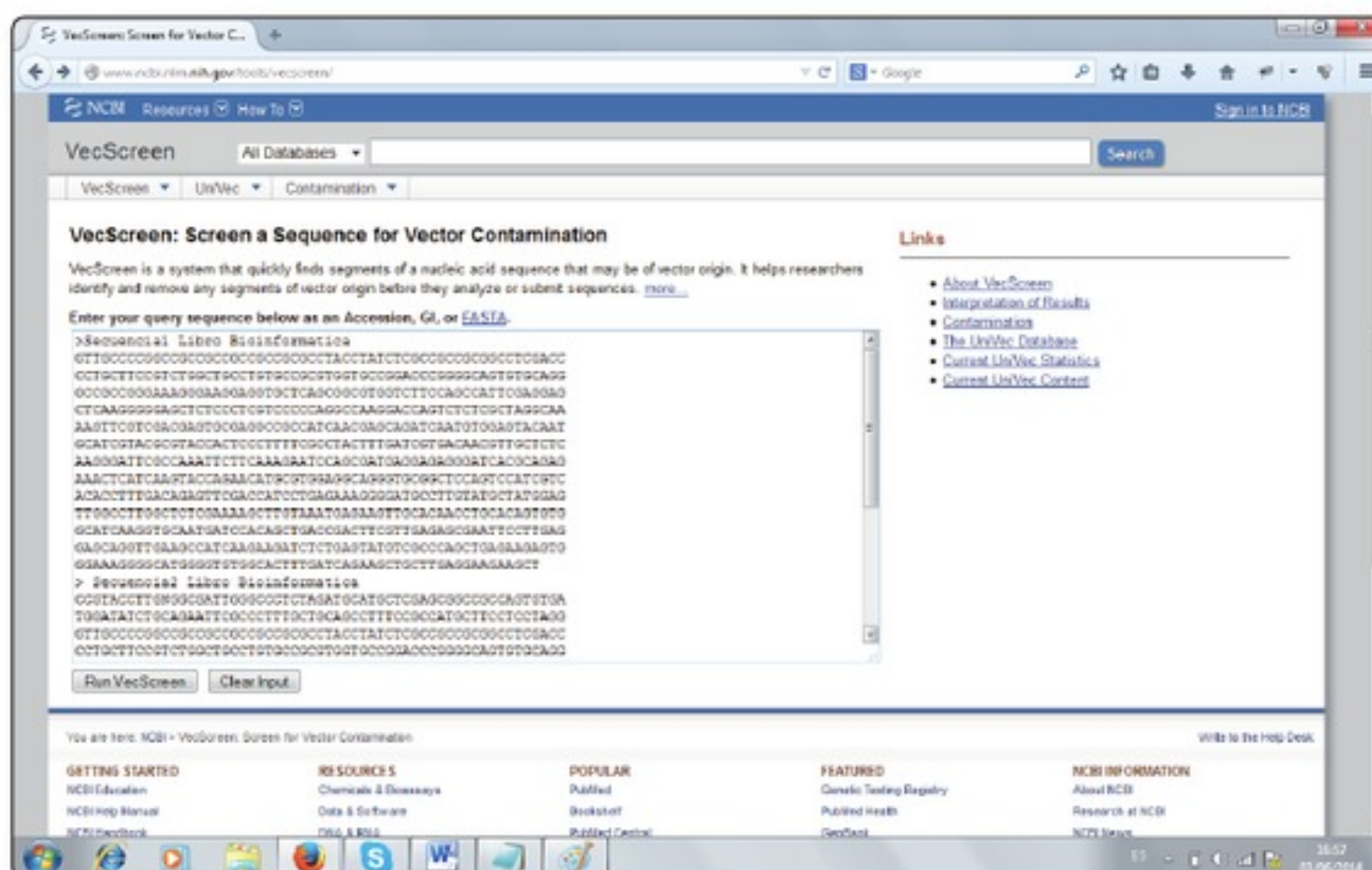


Figura 9.2. Página principal de VecScreen

Finalmente, hacemos clic en el botón **run VecScreen** para obtener la página de resultados, en donde indicamos los parámetros del análisis que se va a realizar y pulsamos en el botón **View Report** (ver Figura 9.3).

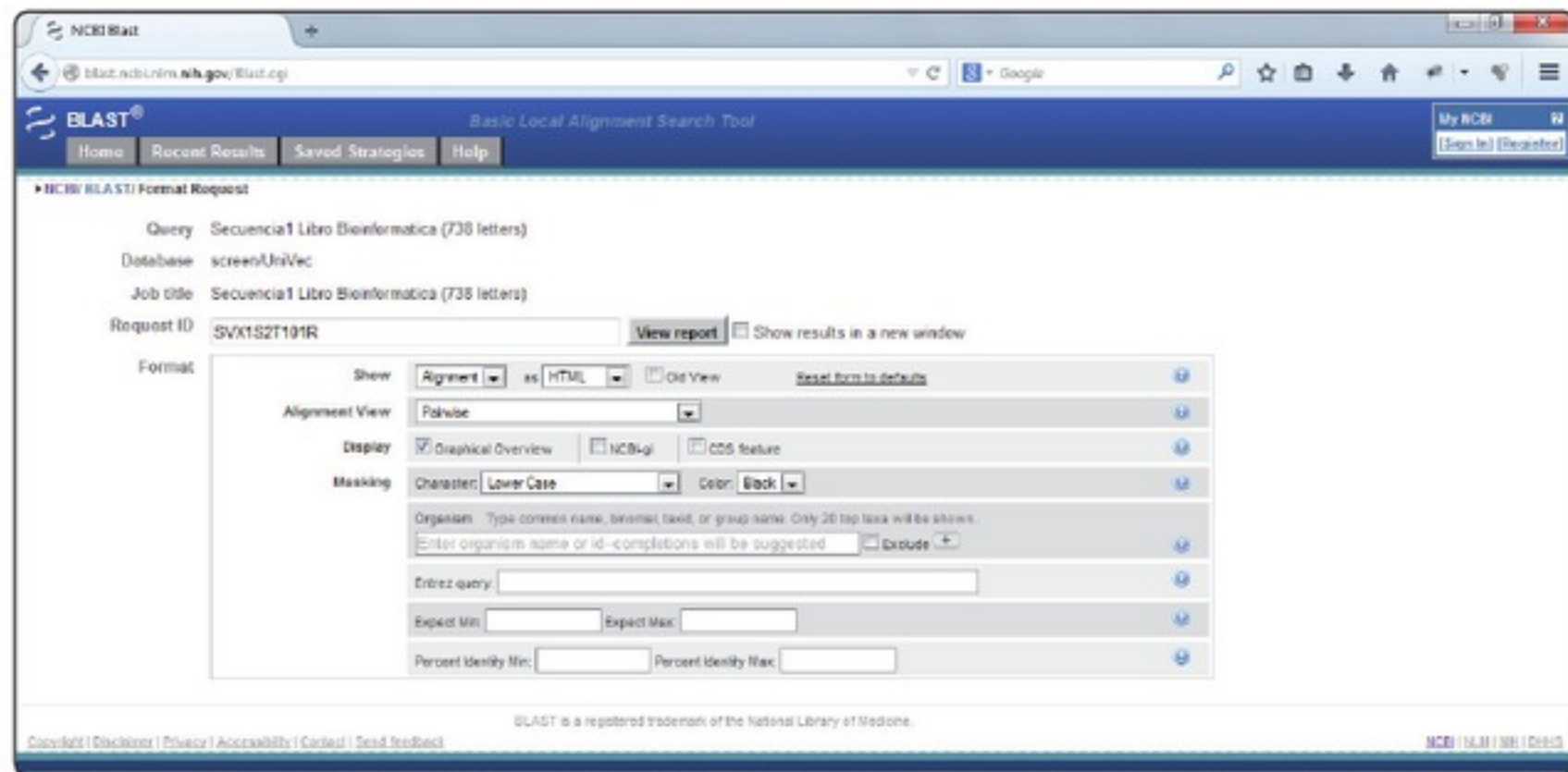


Figura 9.3. Informe de resultados

En la ventana de informe, seleccionamos la primera secuencia (ver Figura 9.4).

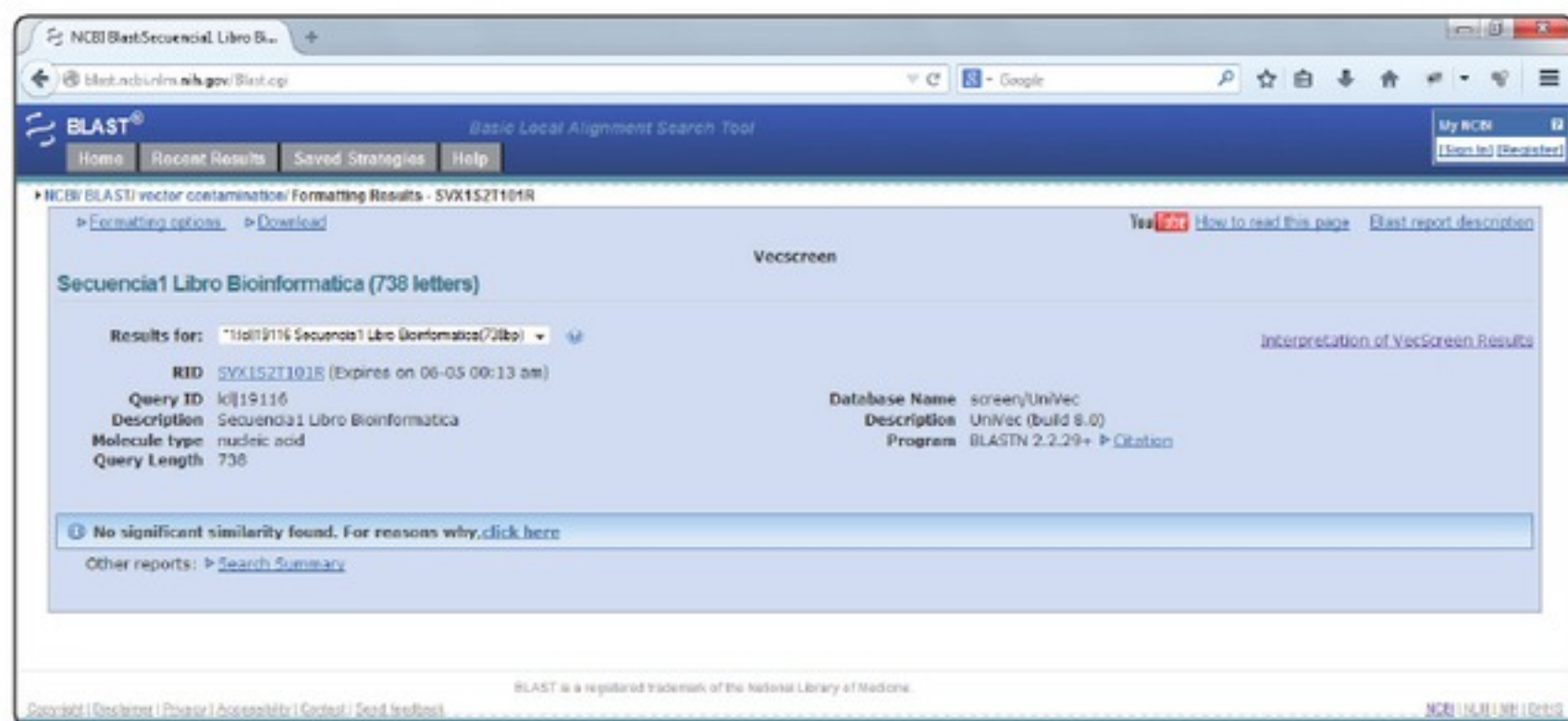


Figura 9.4. Detalles del informe para la primera secuencia

Como puede verse, para esta secuencia VecScreen determina que no ha encontrado ningún resultado, lo que quiere decir que la secuencia no está contaminada y podemos proseguir el análisis de la misma.

Ahora, seleccionamos la segunda secuencia. En este caso, la secuencia sí está contaminada y VecScreen proporciona información útil sobre dicha contaminación.

En la parte superior del informe se muestra la distribución de las secuencias contaminantes con un código de colores similar al de la Figura 9.5:

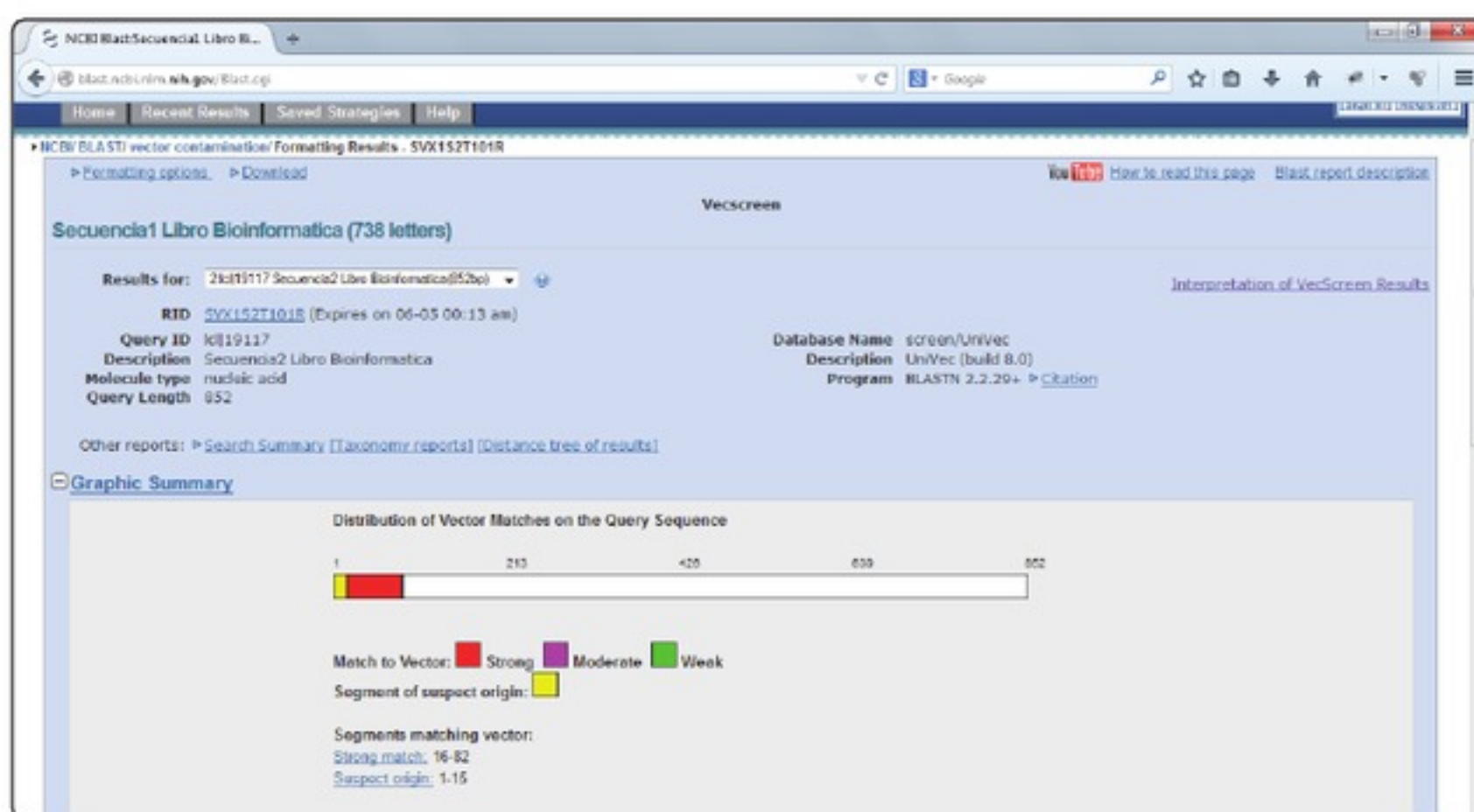


Figura 9.5. Informe de contaminación de la segunda secuencia

Tal y como indica la leyenda, el rojo se corresponde con un alto nivel de contaminación; el púrpura, con un nivel moderado; y, el verde, con un nivel bajo. Un nivel alto o moderado indica que el segmento original de ADN foráneo se adjuntó al ADN/ARN origen durante el proceso de clonación. Por el contrario, un nivel de contaminación débil significa que los segmentos de la secuencia son potencialmente idénticos a los de la secuencia origen.

En la sección de alineamientos (**Alignments**) encontramos los vectores que han contaminado la secuencia bajo estudio (ver Figura 9.6).

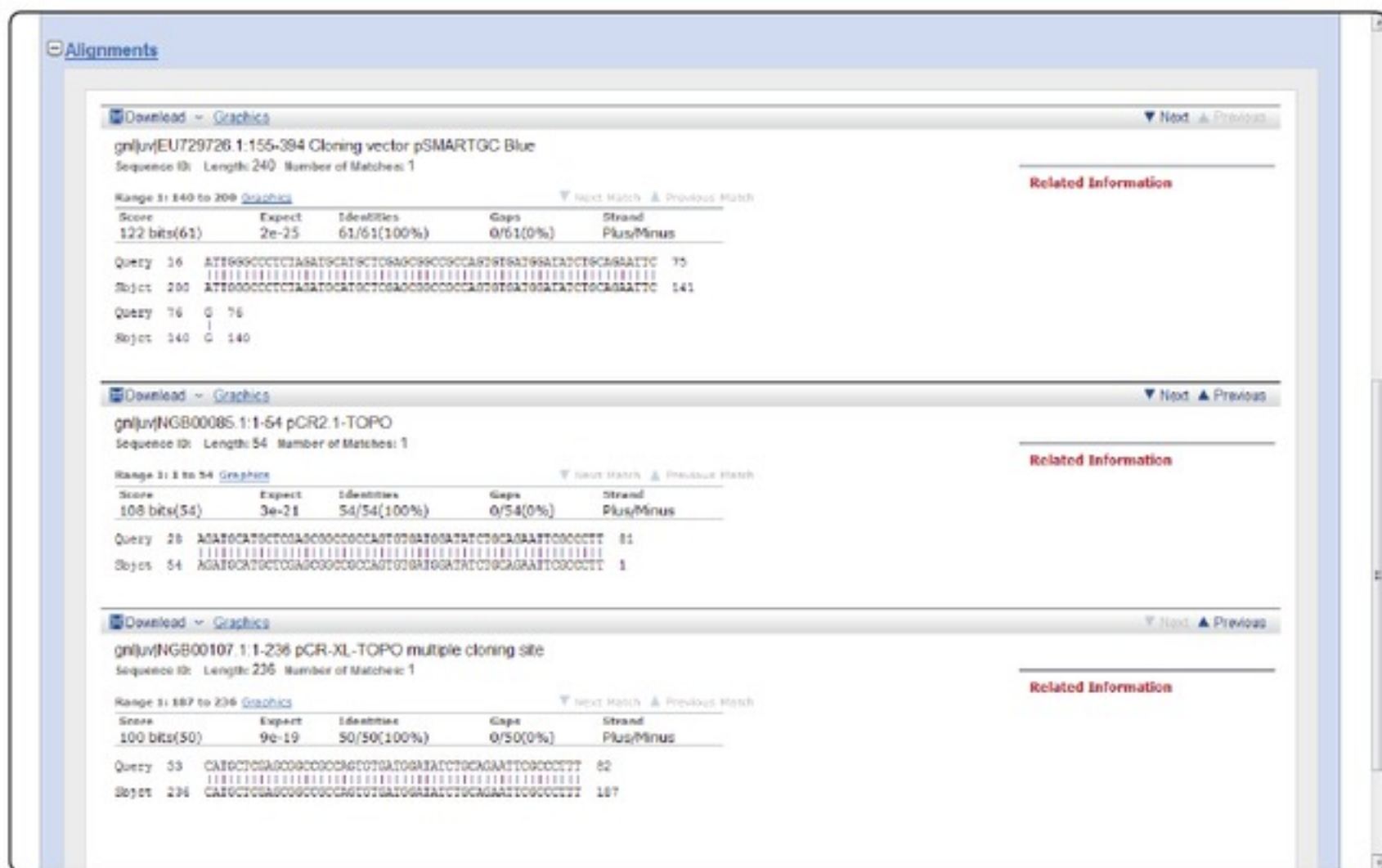


Figura 9.6. Vectores contaminantes

Los segmentos extraños ubicados cerca del final de la secuencia deben eliminarse, excepto las colas PolyA porque suponen una marca fácil de identificar. Una cola PolyA es una cadena de adeninas que se añade al ARNm durante su procesamiento para aumentar la estabilidad de la molécula. Por tanto, eliminaremos cualquier fragmento de secuencia a partir de la posición de la cola PolyA.

Un fragmento de secuencia extraña en medio de la secuencia bajo análisis suele indicar que dos fragmentos de la secuencia nativa se han unido, ya sea durante la clonación o durante el ensamblado de la secuencia. En la mayoría de los casos, el fragmento de secuencia extraña debe eliminarse y dividir la secuencia bajo análisis en dos secuencias separadas.

Basándonos en los resultados de la sección de Alineamientos, se observa que existen secuencias de tres vectores: pSMARTGC Blue, pCR2.1-TOPO y pCR-XL-TOPO *multiple cloning site*. Supongamos que utilizamos el primero durante la clonación.

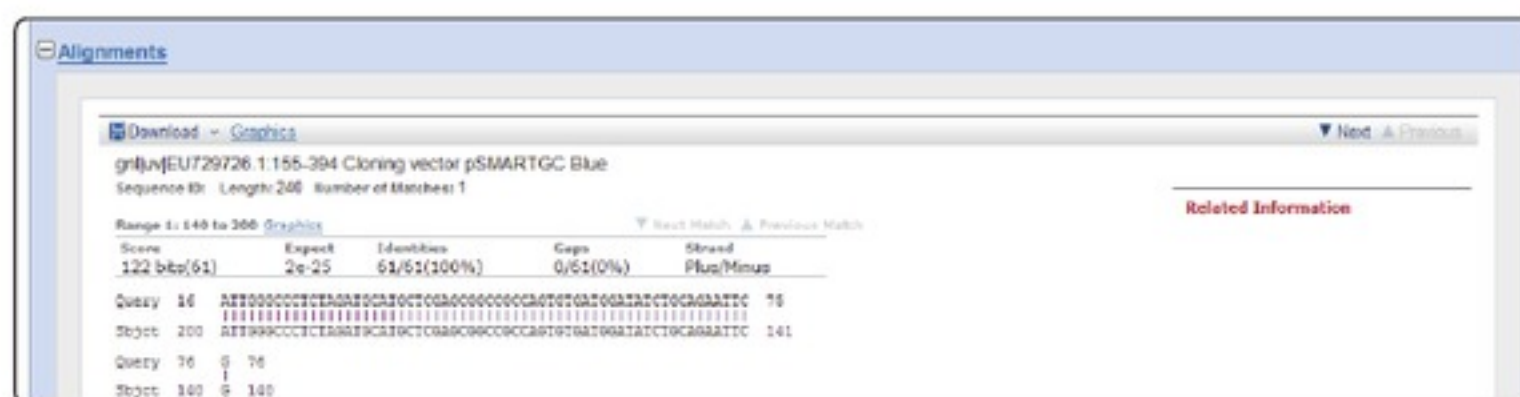


Figura 9.7. Primer vector de clonación

En este caso, se ve que el alineamiento se corresponde con las bases 17-75 de la secuencia bajo estudio y que, por tanto, las bases 1-75 deben eliminarse. Si procedemos así, la secuencia origen, una vez recortada quedará:

```
> Secuencia2 Libro Bioinformatica
GCCCTTTGCTGCAGCCTTTCCGCCATGCTTCCTCCTAGGGTTGCCCGGCCGCGCCGCGCGCGCGC
GCCTACCTATCTCGCCGCGCGGCCCTCGACCCTGCTTCCGTCTGGCTGCCTGTGCCGCGTGGTGC
CGGACCCGGGGCAGTGTGCAGGGCCGCGGGGAAAGGGAAGGAGGTGCTCAGCGGCGTGGTCTTCC
AGCCATTCGAGGAGCTCAAGGGGGAGCTCTCCCTCGTCCCCCAGGCCAAGGACCAGTCTCTCGCT
AGGCAAAAGTTCGTGACGAGTGCGAGGCCGCCATCAACGAGCAGATCAATGTGGAGTACAATGC
ATCGTACGCGTACCACTCCCTTTTCGCCTACTTTGATCGTGACAACGTTGCTCTCAAGGGATTTCG
CCAAATTCTTCAAAGAATCCAGCGATGAGGAGAGGGATCACGCAGAGAACTCATCAAGTACCAG
AACATGCGTGGAGGCAGGGTGC GGCTCCAGTCCATCGTCACACCTTTGACAGAGTTTCGACCATCC
TGAGAAAGGGGATGCCTTGTATGCTATGGAGTTGGCCTTGGCTCTCGAAAAGCTTGTAATGAGA
AGTTGCACAACCTGCACAGTGTGGCATCAAGGTGCAATGATCCACAGCTGACCGACTTCGTTGAG
AGCGAATTCCTTGAGGAGCAGGTTGAAGCCATCAAGAAGATCTCTGAGTATGTCGCCCAGCTGAG
AAGAGTGGGAAAGGGGCATGGGGTGTGGCACTTTGATCAGAAGCTGCTTGAGGAAGAAGCT
```

Figura 9.8. Segunda secuencia recortada

Si ahora se vuelve a ejecutar el análisis de la secuencia (pasos 1 al 5), se obtiene que no se han encontrado similitudes y que, por tanto, la secuencia está “limpia” y podemos proseguir con el análisis.

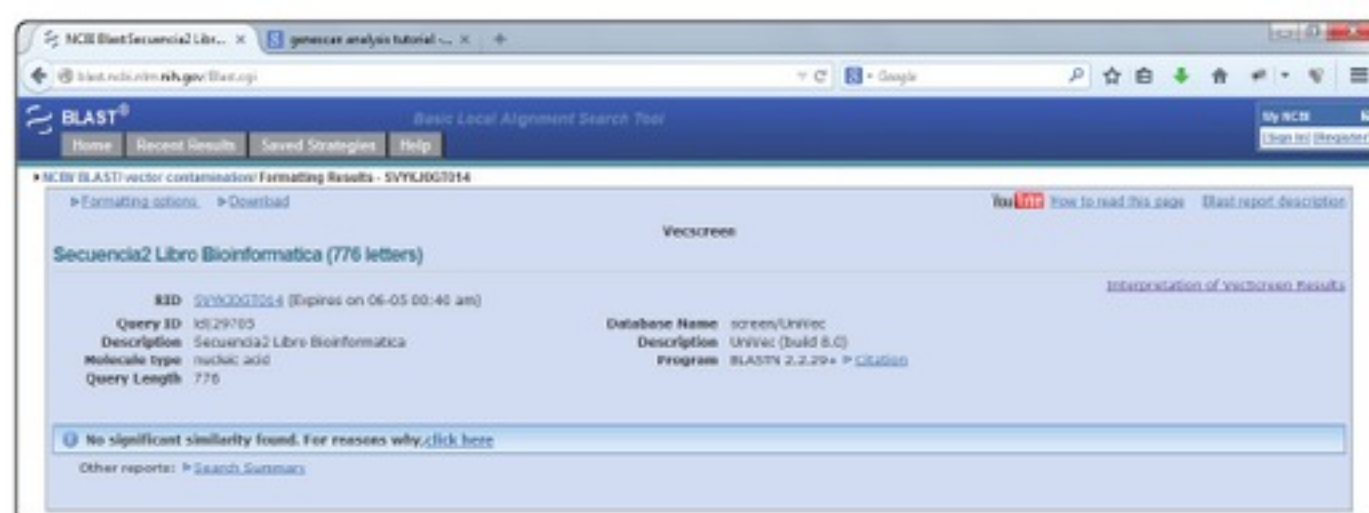


Figura 9.9. Resultado de analizar la secuencia recortada



desechar la secuencia.

En general, si la contaminación se halla en los extremos de la secuencia y se corresponde con el vector empleado, la secuencia se puede limpiar fácilmente y el proceso acaba aquí. Si, por el contrario, la contaminación está dispersa, la mejor opción es

9.2 ANÁLISIS DE LA COMPOSICIÓN DEL ADN

Una vez que estamos seguros de que la secuencia es correcta, ya se está en condiciones de comenzar a obtener información sobre la misma.

9.2.1 Búsqueda de palabras

El Instituto Pasteur proporciona la herramienta WordCount (<http://mobyli.pasteur.fr/cgi-bin/portal.py?#forms::wordcount>) para la búsqueda de palabras, con sentido biológico, en una secuencia dada.

En primer lugar, copiamos la secuencia bajo estudio en el área de texto destinada al efecto (ver Figura 9.10 y Figura 9.11):

```
>Secuencial Libro Bioinformatica
GTTGCCCCGGCCGCCGCCGCCGCCGCGCTACCTATCTCGCCGCCGCCGCCCTCGACCCCTGCTTC
CGTCTGGCTGCCTGTGCCGCGTGGTGCCGGACCCGGGGCAGTGTGCAGGGCCGCCGGGAAAGGGA
AGGAGGTGCTCAGCGGCGTGGTCTTCCAGCCATTGAGGAGCTCAAGGGGGAGCTCTCCCTCGTC
CCCCAGGCCAAGGACCAGTCTCTCGCTAGGCAAAAGTTGTCGACGAGTGCGAGGCCGCCATCAA
CGAGCAGATCAATGTGGAGTACAATGCATCGTACGCGTACCACTCCCTTTTCGCCTACTTTGATC
GTGACAACGTTGCTCTCAAGGGATTGCGCAAATTCTTCAAAGAATCCAGCGATGAGGAGAGGGAT
CACGCAGAGAACTCATCAAGTACCAGAACATGCGTGAGGAGGAGGTTGCGGCTCCAGTCCATCGT
CACACCTTTGACAGAGTTGACCATCCTGAGAAAGGGGATGCCTTGTATGCTATGGAGTTGGCCT
TGGCTCTCGAAAAGCTTGTAATGAGAAGTTGCACAACCTGCACAGTGTGGCATCAAGGTGCAAT
GATCCACAGCTGACCGACTTCGTTGAGAGCGAATTCCTTGAGGAGCAGGTTGAAGCCATCAAGAA
GATCTCTGAGTATGTCGCCCGAGCTGAGAAGAGTGGGAAAGGGGCATGGGGTGTGGCACTTTGATC
AGAAGCTGCTTGAGGAAGAAGCT
```

Figura 9.10. Secuencia bajo estudio

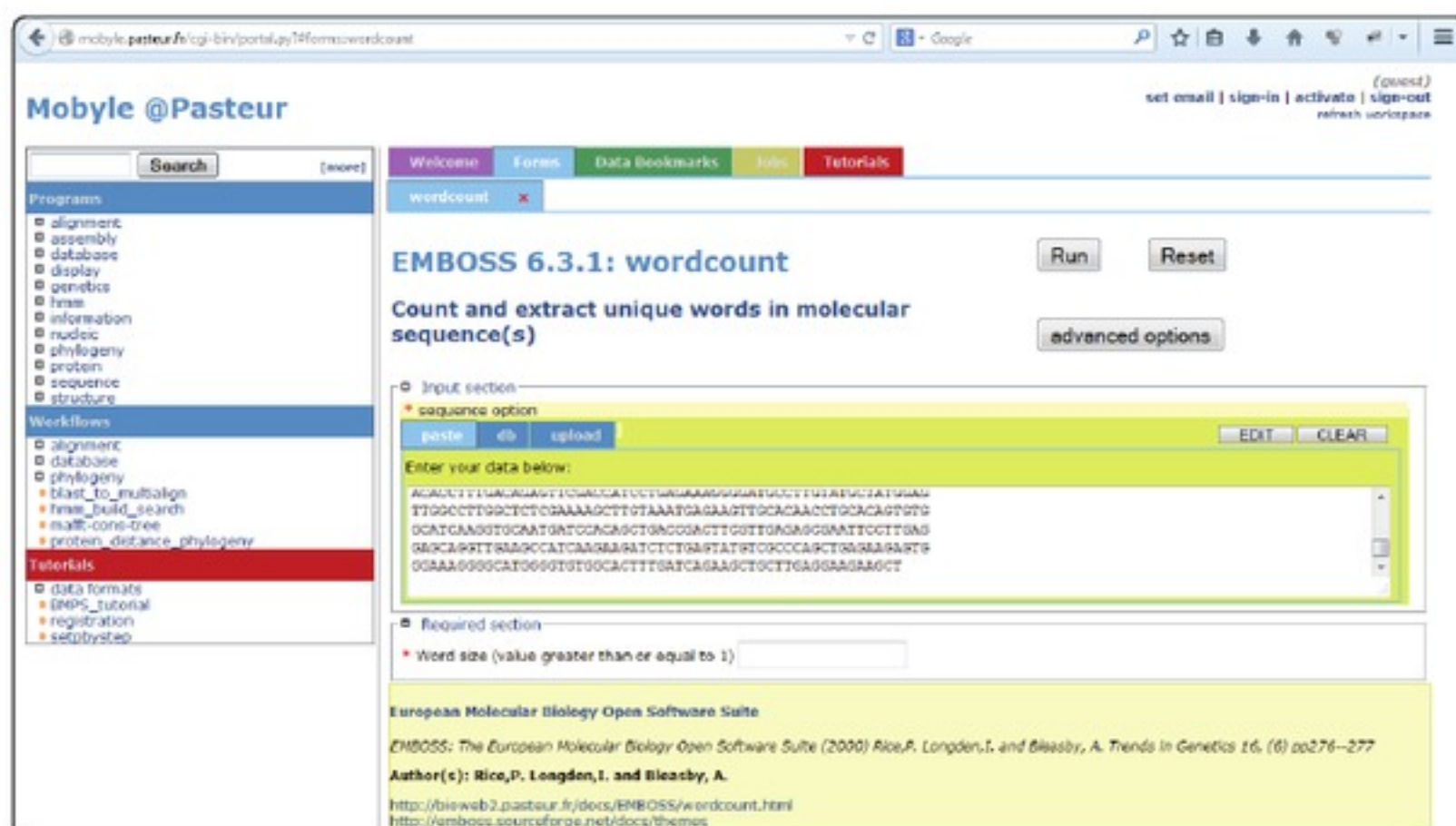


Figura 9.11. WordCount de EMBOSS

Si hacemos clic en **Run** obtenemos la página de resultados (ver Figura 9.12):

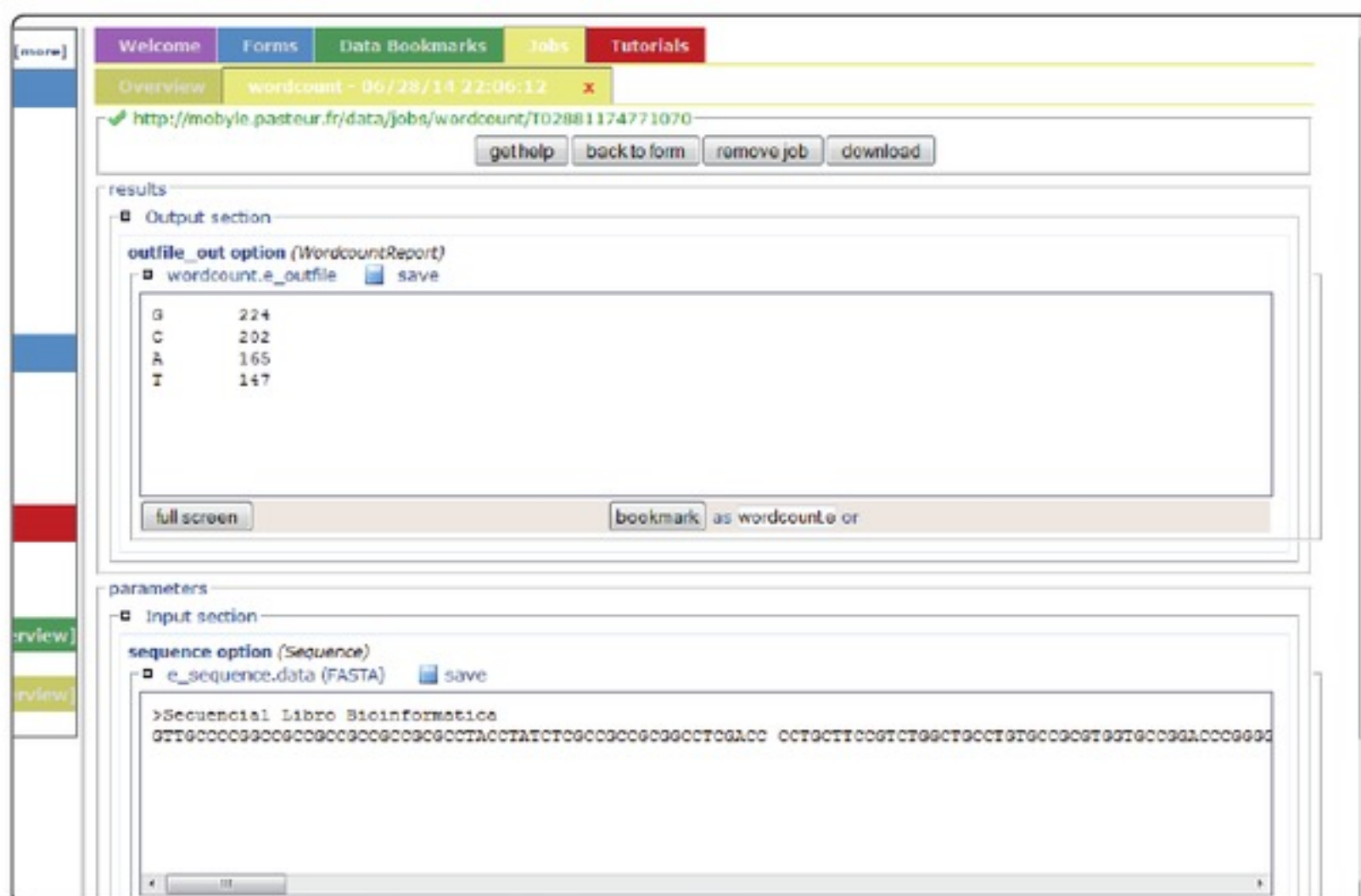


Figura 9.12. Página de resultados

Como vemos, la secuencia bajo estudio está compuesta por 224 guaninas, 202 citosinas, 165 adeninas y 147 timinas. Si ahora, en lugar de contar palabras de 1 nucleótido, se especifican palabras de 3 nucleótidos, los resultados son los de la Figura 9.13. De esta manera, es posible comprobar los aminoácidos que componen la secuencia.

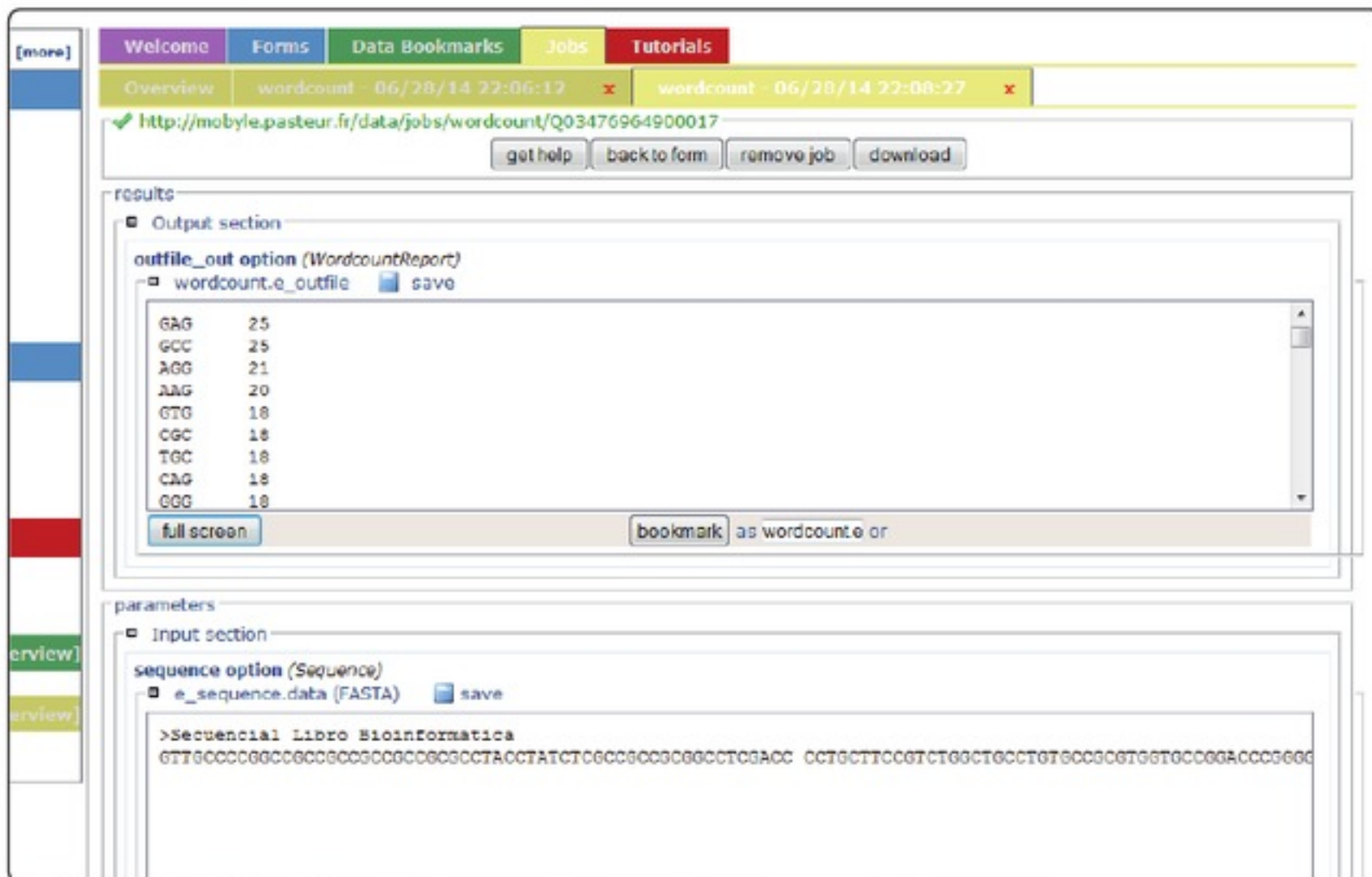


Figura 9.13. Página de resultados con palabras de 3 nucleótidos

9.2.2 Estadísticas de la secuencia con Genomatix

La empresa Genomatix ofrece una herramienta web que permite extraer estadísticas básicas de una secuencia de ADN.

En primer lugar, indicamos la secuencia en la página principal de Genomatix (<http://www.genomatix.de/cgi-bin/tools/tools.pl>) y pinchamos en el botón **Load Sequence**:

The screenshot shows the Genomatix software suite web interface. The main heading is 'Create sequence statistics'. Below this, there are three input options for sequence data:

- Sequence Input:** A text area where a DNA sequence can be pasted. A message says 'Please enter a correctly formatted DNA sequence(s)'. A list of supported formats is shown: plain, DDBL, FASTA, GCG937, GenBank, etc. A sample sequence is provided:


```
GAATTCCTTGAG
GAGCAGTTGAAGCCATCAAGAGATCTGTGATATGTCGCCCG
CTGAGAGAGTG
GGAAAGGGGCGATGGGGTGTGGCATTTCATCAGAGCTGCTTGA
GAAAGAGCT
```
- or enter a file with sequence(s):** A button labeled 'Examinar...' and a message 'No se ha seleccionado ningún archivo. (máx. 100 KB)'. Below it is a field for 'Optional name for your sequence-file on the server:'.
- or enter accession number(s):** A field for entering accession numbers, with a note '(separated by spaces or commas)'.

At the bottom left are buttons for 'Load Sequence' and 'Reset Form'. At the bottom right is a link to 'Register here for a free Genomatix trial account!'.

Figura 9.14. Página principal de Genomatix

Una vez cargada la secuencia, pinchamos en el botón **Start Task** para iniciar el procesamiento de la secuencia. Finalmente, obtendremos unos resultados como los de la Figura 9.15:

The screenshot shows the results page after uploading a sequence. The heading is 'Statistics for all input sequences (738 basepairs in 1 sequences):'. Below this, there are three sections: 'Base-Content:', 'Mono-Nucleotides:', and 'Di-Nucleotides:'.

Base-Content:

The AT-content of this sequence is 42.28%, the GC-content 57.72%, others: 0.00%

Mono-Nucleotides:

	in bp	in %
A	165	22.50%
C	262	27.37%
G	224	30.35%
T	147	19.92%

Di-Nucleotides:

		second nucleotide			
		A	C	G	T
first nucl.	A	43	28	64	30
	C	51	57	45	48
	G	59	66	58	41
	T	12	51	53	30

Tri-Nucleotides:

		second/third nucleotide															
		AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
first nucl.	A	11	5	20	7	8	10	5	5	16	14	21	13	0	15	11	4
	C	15	8	18	10	16	11	16	14	11	18	7	12	5	17	10	13
	G	16	9	25	9	14	25	11	16	17	16	18	7	7	0	10	8
	T	1	6	1	4	13	11	16	11	15	18	12	8	0	11	14	5

Figura 9.15. Estadísticas básicas de la secuencia bajo estudio

9.2.3 Búsqueda de repeticiones

Una repetición es un segmento de secuencia que ocurre más de una vez en la secuencia bajo estudio y que es suficientemente larga como para que la repetición no sea aleatoria. El interés en las repeticiones se debe a que suelen estar implicadas en mecanismos de regulación de la expresión del gen.

La principal diferencia entre el conteo de palabras y el análisis de repeticiones es que no es necesario que las repeticiones sean idénticas. De hecho, se distingue entre repeticiones en tándem (fragmentos de secuencia similares en la misma hebra de ADN) y repeticiones invertidas (fragmentos de secuencia similares que ocurren en la hebra directa y la inversa).

A la hora de analizar las repeticiones, el punto crítico es el umbral de repetición, que es el número de nucleótidos iguales a partir de los cuales se puede considerar que dos fragmentos de una secuencia constituyen una repetición.



Los algoritmos de búsqueda de repeticiones dependen de gran cantidad de parámetros, por lo que se recomienda contrastar los valores con distintas herramientas y, sobre todo, no dar por definitiva una búsqueda de repeticiones infructuosa.

No hay que confundir la identificación de una repetición en la secuencia bajo estudio con la identificación de una repetición registrada en una lista de repeticiones predefinidas. En el primer caso, el descubrimiento está relacionado con la estructura interna de la secuencia bajo estudio, mientras que en el segundo hace referencia al reconocimiento de cierta similitud entre la secuencia bajo estudio y un conjunto predefinido de repeticiones almacenadas en una base de datos como RepBase del Genetic Information Research Institute (<http://www.girinst.org>).

Una manera de contrastar los resultados ofrecidos por distintas herramientas es utilizar un gráfico de puntos, sobre todo, cuando estamos trabajando con secuencias de ADN de una longitud considerable. Para aprender cómo utilizar los gráficos de puntos, emplearemos la siguiente secuencia:

```
>Secuencial Libro Bioinformatica
GTTGCCCCGGCCGCCGCGCCGCGCGCTACCTATCTCGCCGCGCGGCCTCGACCCCTGCTTC
CGTCTGGCTGCCTGTGCCGCGTGGTGCCGGACCCGGGGCAGTGTGCAGGGCCGCCGGGAAAGGGA
AGGAGGTGCTCAGCGGCGTGGTCTTCCAGCCATTCGAGGAGCTCAAGGGGGAGCTCTCCCTCGTC
CCCCAGGCCAAGGACCAGTCTCTCGCTAGGCAAAGTTCGTGACGAGTGCGAGGCCGCCATCAA
```



```
CGAGCAGATCAATGTGGAGTACAATGCATCGTACGCGTACCACTCCCTTTTCGCCTACTTTGATC
GTGACAACGTTGCTCTCAAGGGATTTCGCCAAATTCTTCAAAGAATCCAGCGATGAGGAGAGGGAT
CACGCAGAGAACTCATCAAGTACCAGAACATGCGTGGAGGCAGGGTGCGGCTCCAGTCCATCGT
CACACCTTTGACAGAGTTCGACCATCCTGAGAAAGGGGATGCCTTGTATGCTATGGAGTTGGCCT
TGGCTCTCGAAAAGCTTGTAATGAGAAGTTGCACAACCTGCACAGTGTGGCATCAAGGTGCAAT
GATCCACAGCTGACCGACTTCGTTGAGAGCGAATTCCTTGAGGAGCAGGTTGAAGCCATCAAGAA
GATCTCTGAGTATGTCGCCCAGCTGAGAAGAGTGGGAAAGGGGCATGGGGTGTGGCACTTTGATC
AGAAGCTGCTTGAGGAAGAAGCT
```

Una de las herramientas más utilizadas para dibujar gráficos de puntos es el Dotlet del SIB (<http://myhits.isb-sib.ch/cgi-bin/dotlet>). Este sitio web es muy útil para el procesamiento simple de secuencias de ADN.

En la página principal de SIB-Dotlet (ver Figura 9.16), introduciremos la secuencia en la ventana que se abre al pinchar el botón **input**, tras lo cual haremos clic en el botón **Ok** de dicha ventana. Una vez que se ha introducido la secuencia, ya no es posible realizar ninguna modificación en la misma.

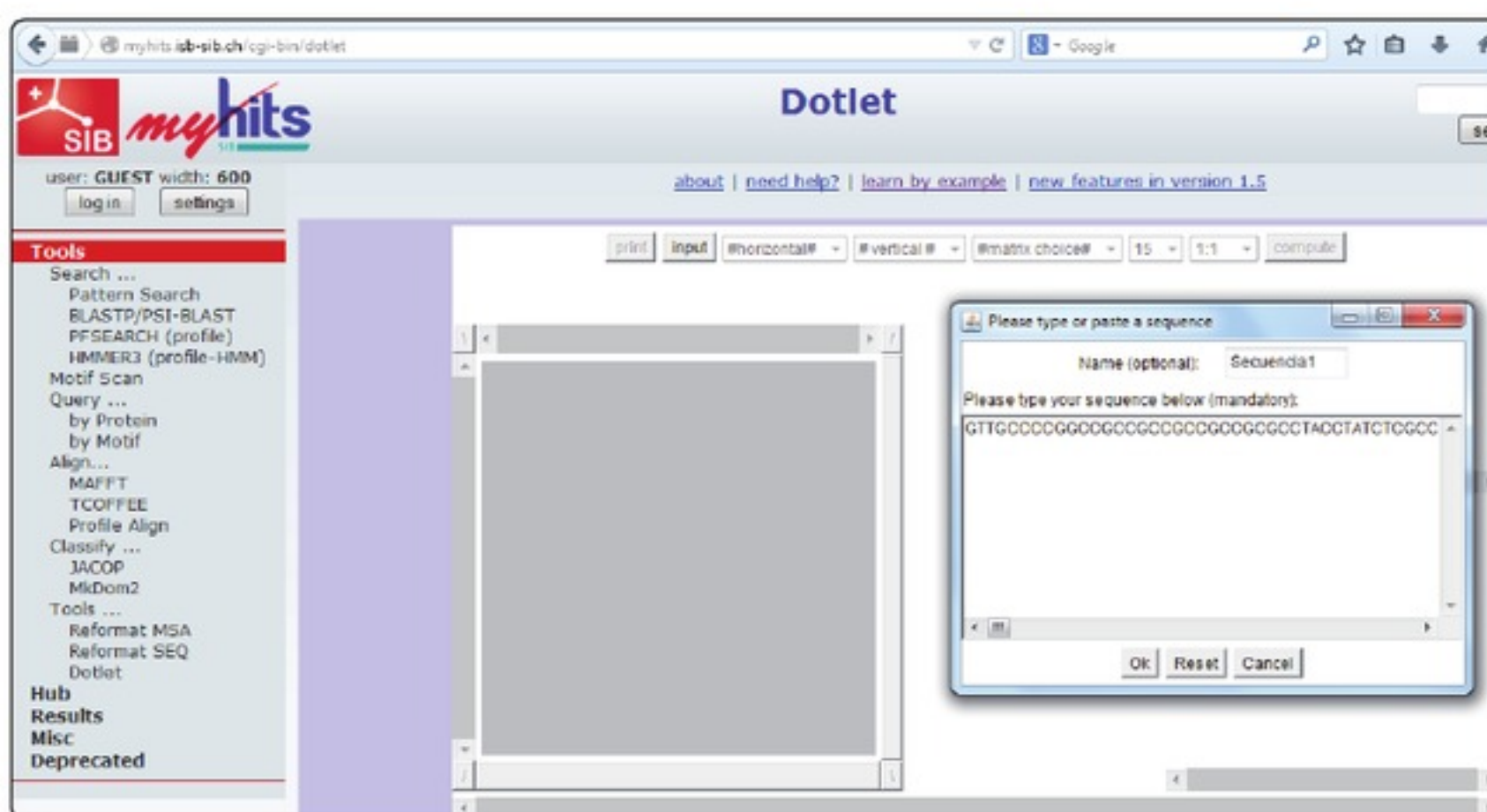


Figura 9.16. Introducción de una secuencia en Dotlet



Al pegar la secuencia en la ventana, no hay que incluir el nombre de la misma, sino la secuencia de nucleótidos en texto plano.

barra de desplazamiento bajo la ventana del histograma hasta una posición en la que las similitudes bajas (las más abundantes) se filtren, esto es, hasta frecuencias del histograma relativamente pequeñas. Llegados a este punto, el gráfico se oscurecerá, tal y como muestra la Figura 9.18.

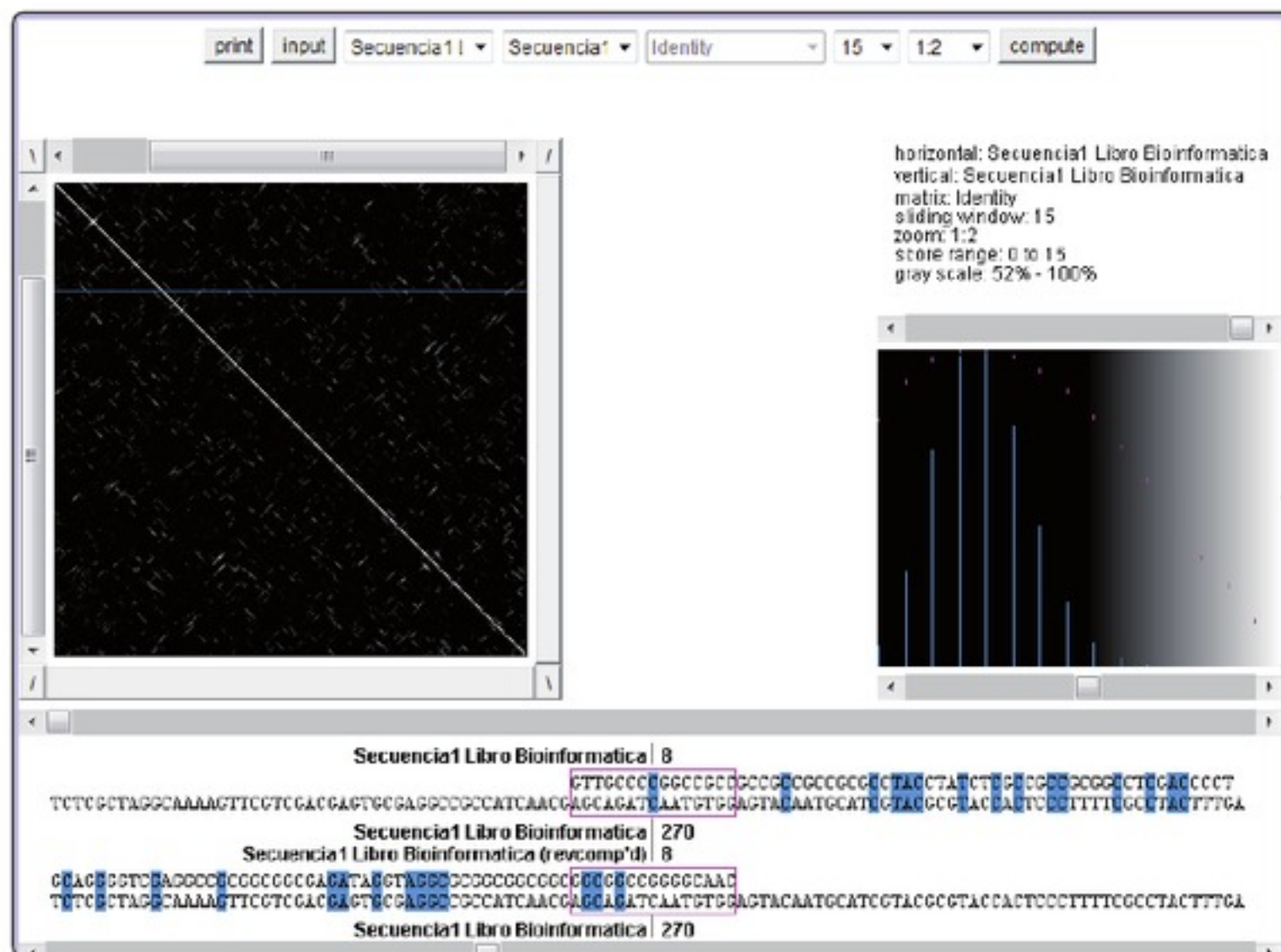


Figura 9.18. Filtrado del ruido de fondo

Una vez filtrado el ruido de fondo, podemos identificar las repeticiones como las líneas que aparezcan a ambos lados de la diagonal. Si las líneas son perpendiculares, están asociadas a secuencias palindrómicas o repeticiones invertidas.



Una secuencia palindrómica es aquella que se lee igual hacia delante que hacia atrás. Por ejemplo, AGGA.

Si cambiamos el tamaño de la ventana a 11 y redibujamos el gráfico de puntos, aparecen áreas con una elevada densidad de puntos, como los rectángulos de

la parte superior de la Figura 9.19. Estas áreas son secuencias repetidas con pocos aminoácidos, como se puede comprobar en la sección de alineamiento, de la que hablaremos más adelante.

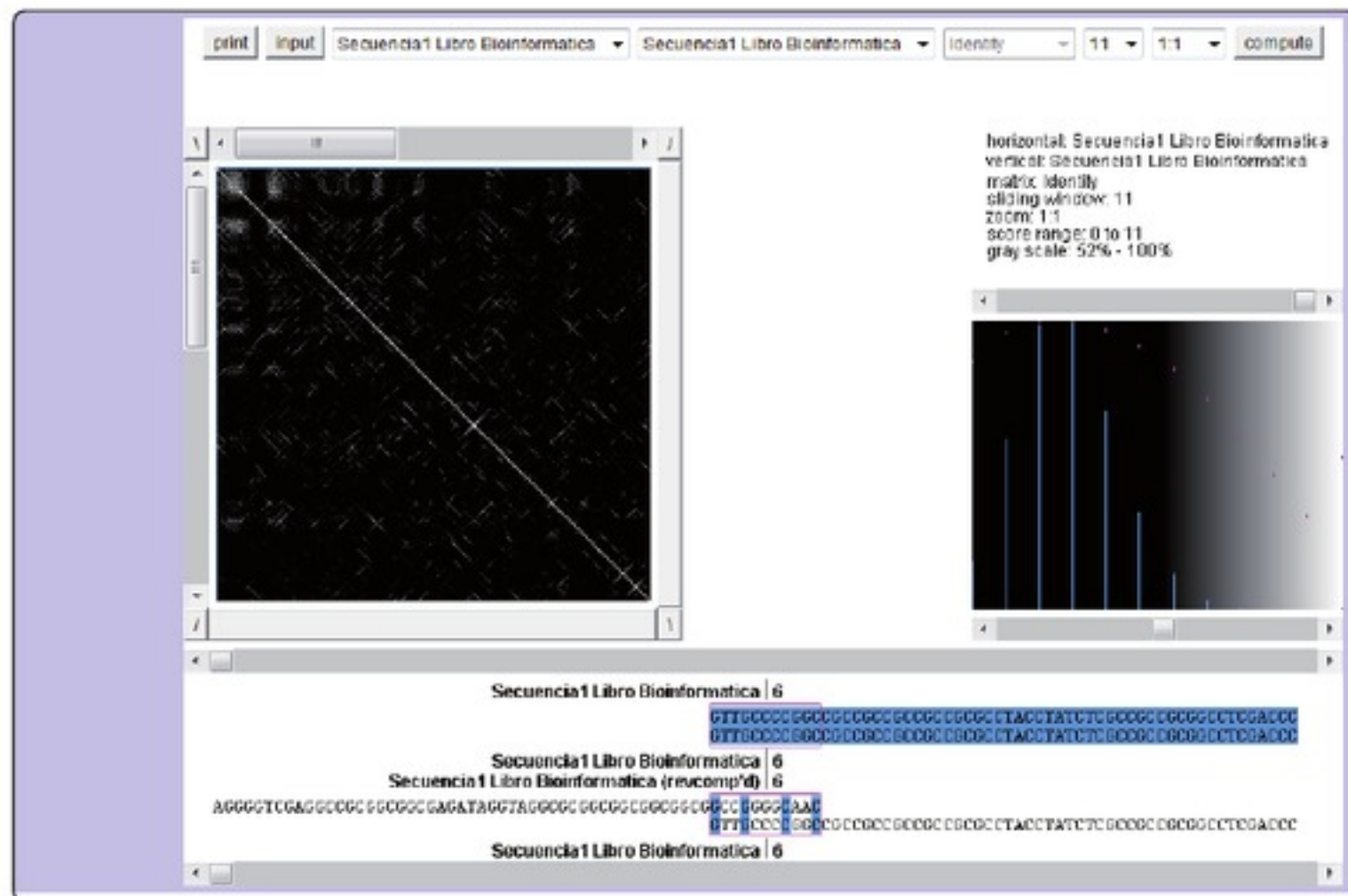


Figura 9.19. Recálculo del gráfico de puntos

El alineamiento de la secuencia bajo estudio consigo misma puede verse en la parte inferior de la pantalla (ver Figura 9.20). Los residuos que están en azul indican que son idénticos y la ventana de comparación está marcada en violeta. Moviendo las barras de desplazamiento, la ventana se va desplazando a lo largo de las secuencias.

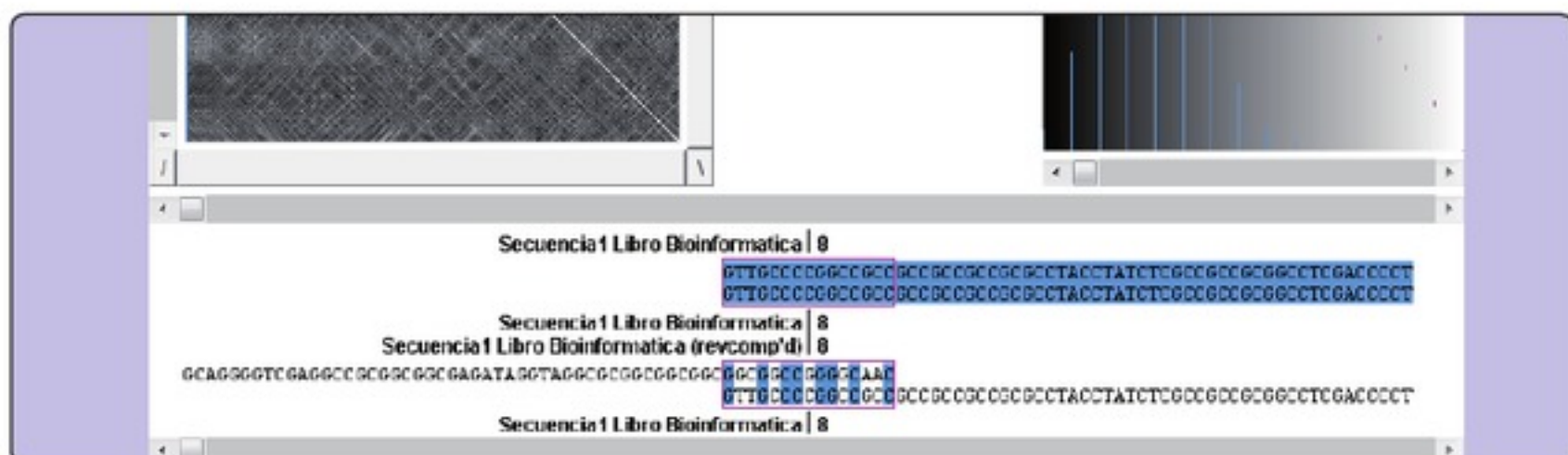


Figura 9.20. Sección de alineamiento

Veamos ahora qué ocurre cuando comparamos dos secuencias diferentes y como el gráfico de puntos puede ayudar en la identificación de secuencias. Para ello, trabajaremos con las siguientes dos secuencias, correspondientes al ARN mensajero de Cfr de ratón (*accession* NM_021050.2) y su equivalente humano (NM_000492.3).

```
>gi|116008179|ref|NM_021050.2| Mus musculus cystic fibrosis
transmembrane conductance regulator (Cftr), mRNA
AATTGGAAGCAAATGACATCACCTCAGGTCTGAGTAAAAGGGACGAGCCAAAAGCATTGACCTGG
TCCTGGATATCCAGATGTCGAGTCCAACCTGAATTTAGCCGAACACAGACCTCATTGCCTCACGG
AGACATCATGCAGAAGTCGCCTTTGGAGAAAGCCAGCTTTATCTCCAAACTCTTCTTCAGCTGGA
CCACACCAATTTTGAGGAAAGGGTACAGACACCACTTGGAGTTGTCAGACATATAACCAAGCCCCT
TCTGCTGATTGAGCTGACCACTTGTCTGAAAACTAGAAAGAGAATGGGACAGAGAACAAGCTTC
AAAAAAGAATCCCCAGCTTATCCACGCCCTTCGGCGATGCTTTTTCTGGAGATTCCTCTTCTATG
GAATTTTGCTATACCTAGGGGAAGTCACCAAGGCTGTCCAGCCTGTCTTGCTAGGAAGAATCATA
GCATCCTATGATCCAGAAAACAAGGTGGAACGTTCCATTGCCATTTACCTTGGCATAGGCTTATG
CCTTCTCTTCATTGTCAGGACACTGCTTCTTCACCCAGCTATTTTTGGCCTTCATCGCATTGGAA
TGCAGATGAGAACAGCTATGTTTAGCTTGATTTATAAGAAGACTTTAAAGTTGTCAAGCCGCGTT
CTTGATAAAATAAGTATTGGACAACCTTGTTAGTCTTCTTTCCAACAACCTGAACAAATTTGATGA
AGGACTTGCCCTGGGCACATTTTATATGGATTGCTCCTTTACAAGTGAAGTCTTCTGATGGGGCTTC
TCTGGGACTTGTTACAGTTCTCAGCCTTCTGTGGCCTTGGTTTACTGATAATCCTGGTTATTTTT
CAAGCTATCCTAGGGAAGATGATGGTGAAGTACAGAGATCAGAGAGCTGCAAAGATCAATGAAAG
ACTCGTGATCACATCAGAAATTATTGATAATATCTATTCTGTTAAGGCATATTGTTGGGAATCAG
CGATGGAGAAAATGATTGAAAACCTGAGAGAGGTGGAGCTGAAAATGACCCGGAAGGCGGCCTAT
ATGAGGTTCTTCACTAGCTCTGCCTTCTTCTTTTCAGGGTTCTTTGTAGTCTTTCTATCTGTGCT
TCCCTACACAGTCATCAACGGAATCGTCCTACGAAAAATATTCACAACCATTTTATTCTGCATTG
TCCTACGTATGTCAGTCACACGGCAGTTCCCCACTGCCGTACAGATATGGTATGATTCTTTTGG
ATGATAAGAAAAATACAGGATTTCTTGCAGAAAACAAGAGTATAAAGTACTGGAGTATAACTTAAT
GACCACAGGCATAATCATGGAAAATGTAACAGCATTTTGGGAGGAGGGATTTGGGGAATTACTGG
AGAAAGTACAACAAGCAATGGTGACAGAAAACATTCCAGTGATGAGAACAAATGTCAGTTTCAGT
CATCTCTGCCTTGTGGGAAATCCTGTGCTGAAAAACATCAATTTGAATATAGAGAAAGGAGAGAT
GTTGGCTATTACTGGATCTACTGGATCAGGAAAGACATCACTCCTGATGTTGATTTTGGGAGAAC
TGGAAGCTTCAGAGGGAATTATTAAGCACAGTGGAAGAGTTTCATTCTGCTCTCAATTTTCTTGG
ATTATGCCGGGTACTATCAAAGAAAAATATCATCTTTGGTGTTCCTATGATGAGTACAGATATAA
GAGTGTTGTCAAAGCTTGCCAACTACAGCAGGACATCACCAAGTTTGCAGAACAAGACAACACAG
TTCTTGGAGAAGGTGGAGTCACACTGAGTGGAGGTGAGCGTGCAAGGATTTCTTTAGCAAGAGCA
GTATATAAAGATGCTGATTTGTACCTATTAGATTCCCCTTTTGGATATCTAGATGTTTTTACTGA
AGAACAAGTATTTGAAAGCTGTGTTTGTAAATTGATGGCCAACAAAACCTAGGATTTTGGTTACAT
CTAAAATGGAACACTTAAGGAAAGCTGACAAAATACTAATTTTGCATCAGGGCAGTAGCTATTTT
TATGGGACATTTTCTGAGCTACAAAGTCTACGTCCAGACTTCAGTTCGAAACTCATGGGGTATGA
TACTTTTGACCAGTTTACTGAGGAAAGAAGAAGTTCAATTCTAACTGAGACCTTACGCAGGTTCT
CAGTAGACGATTCCCTCTGCCCCGTGGAGCAAACCCAAACAGTCGTTTACAGAGACTGGAGAGGTG
GGAGAAAAAAGGAAGAACTCTATTCTAAATTCATTGAGCTCTGTAAGGAAAATTTCCATTGTGCA
AAAGACTCCATTATGTATCGATGGAGAGTCTGATGATCTCCAAGAAAAGAGACTGTCCCTAGTTC
CGGATTCTGAACAGGGGGAGGCTGCTCTGCCGCGCAGCAACATGATCGCCACCGGCCCCACATTT
CCAGGCAGAAGAAGACAGTCTGTTTTGGATCTGATGACGTTACACCCAACTCAGGCTCCAGCAA
TCTTCAGAGGACCAGAACTTCTATTGAAAAATCTCCTTAGTCCCTCAGATAAGCTTAAATGAAG
```


TGGATGTATATTCAAGGAGATTATCGCAAGATAGCACACTGAACATCACTGAAGAAATTAACGAA
GAAGATTTAAAGGAGTGTTTTCTTGATGATGTGATCAAGATACCCCCGGTGACAACATGGAACAC
ATACCTACGATATTTTACTCTCCATAAAGGCTTACTGCTAGTGCTGATTTGGTGCGTACTGGTTT
TTCTGGTTGAGGTGGCTGCTTCTTTATTTGTGTTATGGTTGCTTAAAAACAACCCTGTTAACAGT
GGAAACAATGGTACTAAAATTTCCAATAGCTCCTATGTTGTGATCATCACCAGTACCAGTTTCTA
TTATATTTTTTACATTTACGTGGGAGTGGCTGACACTTTGCTTGCCCTGAGCCTCTTCAGAGGTT
TGCCGCTGGTGACATCGTTAATCACAGCATCAAAAATTTTGACAGGAAAATGTTACACTCCATT
CTTCACGCCCCCTATGTCGACCATCAGCAAGCTGAAAGCAGGTGGGATTCTTAACAGATTCTCCAA
AGATATAGCAATTTTGGATGACTTTCTGCCTCTTACCATTTTTGACTTCATTTCAGTTGGTGTTCA
TTGTGATTGGAGCTATAATAGTCGTCTCGGCATTACAACCCTACATCTTCCTAGCAACGGTGCCA
GGGCTAGTAGTCTTTATTTTACTGAGGGCCTACTTCCTTCATACAGCACAGCAGCTCAAACAAC
GGAATCTGAAGGCAGGAGTCCAATTTTCACCCACCTTGTGACAAGCTTAAAAGGACTCTGGACAC
TTCGAGCCTTCCGACGCCAGACTTACTTTGAAACTCTGTTCCACAAAGCTCTGAATTTGCACACT
GCCAACTGGTTTATGTATCTGGCAACCTTGCGCTGGTTCCAAATGAGAATAGACATGATATTTGT
CCTCTTCTTCATTGTTGTTACCTTCATCTCCATTTTAACAACAGGTGAAGGAGAAGGAACAGCTG
GTATTATTCTAACTTTAGCTATGAATATCATGAGTACTTTGCAGTGGGCTGTGAACCTCAAGCATT
GATACAGATAGCTTGATGCGATCTGTGAGCAGAGTGTTTAAGTTTATTGATATACAAACAGAAGA
AAGTATGTACACACAGATAATTAAAGAACTACCTAGAGAAGGATCATCTGACGTTTTAGTCATTA
AGAATGAGCATGTGAAGAAAAGTGATATCTGGCCCTCTGGAGGCGAAATGGTTGTCAAAGACCTT
ACTGTGAAATACATGGATGATGGAAATGCCGTATTAGAGAACATTTCTTTTTCAATAAGTCCTGG
ACAGAGGGTGGGGCTCTTAGGAAGAACTGGATCAGGAAAAAGTACTTTGCTTTCAGCATTTTTAC
GAATGTTGAACATTAAAGGTGATATAGAGATTGATGGTGTCTCATGGAATTCAGTGACCTTACAA
GAATGGAGGAAAGCTTTCGGAGTGATAACACAGAAAGTATTTATCTTTTCTGGAACATTCAGACA
AAACCTGGATCCCAATGGAAAATGGAAAGATGAAGAAATATGGAAAGTTGCAGATGAGGTTGGAC
TCAAGTCTGTAATAGAGCAGTTTCCTGGACAGCTCAACTTTACCCTTGTGGATGGGGGTTATGTG
CTAAGCCATGGCCATAAGCAATTAATGTGCTTGGCCCGATCAGTTCTCAGTAAGGCCAAGATCAT
ACTGCTTGATGAGCCCAGTGCCCATCTAGACCCCATACATAACCAAGTCATTCGACGAGTTCTAA
AACAAGCCTTCGCTGGTTGCACAGTCATCCTCTGTGAACACAGGATAGAAGCGATGTTGGATTGC
CAGCGATTTTTGGTCATAGAAGAGAGCAATGTCTGGCAGTACGACTCCCTTCAGGCACTTCTGAG
TGAGAAGAGTATCTTCCAGCAGGCCATTAGCTCCTCGGAAAAGATGAGGTTCTTCCAGGGCCGCC
ACTCCAGCAAGCACAAAGCCTCGGACGCAAATTAAGTCTGCTGAAAGAGGAGACAGAAGAAGAAGTT
CAAGAAACCCGTCTCTAGTGCTGGGATGCTGAGGAAGCAACTCAGTGCCTGAGTCCATTCCCAG
AACCCATGCAGAATGAAAAAAGCCAGGCATTTCCCATGCTTCTAACCCAGTGCTGGGGACACAG
AGACAGGTGGATCCCTGGGGCTCTGTGGCAAGTGATCCTAGCCCAACAAAGAGAGTTCCAGGCTGG
GCACCTGAGGGACAATACCTGTGGATATACTCTTGCTTCCACATGCAAGTACATATACACATGCA
TGCACATTAGTGGACATACACACAGAAAAGCAAAGAAGAAGGAAAGAGGGAAGAAAATAGTGCAA
ATAATTGCAAAACGATCATGTATGGAGTCTGCTCATGGACTTAGAGGAGGTGAACTCTACTACCT
GTGCCTTTGAAAGAAGGGTGAAGCCTGCGACTTGCTCTTTAAGAGACTGTTTTGGAAGAGAGTTC
AAAAACGTTTCATATGGGTATGGGTAACTGACTTTCCAGCAGTAGTCAAATTGTTTGAACCTTCAGA
TAGTTGATAATGACCACTTGTGTATTGCAAGGCAGATTTTTCTGAAAACATTTGCCCCCTAATAG
TAGCTGAAAAAGCAGCTATAAATGCCAACCAGGTTAGTCATTCGGCTTATTGTTTCAGTACAGCTG
GTTAATTTGCATTATTGAAGAACTGAAATTATAGTGCTTAGATATAGGACAAAGTAAAGAGAAGT
AAAAACAGTGTCTTATATACTCAAAGCCCAACTTACTTTCTCTAAGATATGTATTGCCTTCTA
TACATTGTCTGCCCCATTCCAAGCAAATGTTAGAATATTATACAAAATACTGGGTGGTATTGATT
GAAAGATGCCCCGACATCTGGTGATCTAGTAACCCATCAGGATTAAGGATATCCAGGTCTTGAAAA
TTAAGGTAAAGACCATCTAGCCTTACTACCGTACAGCTAAACATTCTTATTACCAGAATAAGACC
TAGGAAAAGAACTGTTTCAGTCCCATAAAGTGGCCTGGATAATTTCTTGTATGGAATCGACA

CACTTATGTTCCCAGAAAGCAACAGATCTTTAAGACTTCTGAAGTGAAGGAAGGTTGTGTTAGTG
CAAAC TAGTGCAGCCCAGTGCCAGGTCCAGGAGTTAACATGTAGACAGGCCATGGACTGTGTGGG
TAGATGCTCATGGAAATGTGCAGTAGTATGTTTCATGTGCTCTCAGCTAGCTGTGTGTACTTCAA
CTGTCTCCACAGAGTTGTTGGGGAGACACTCTGAAAAAGAATTAATTGTGAATTAGTTTTATATA
CTTTGTTTTATAATTTGTGATGCAAATGAAAATTTCTCTGGGAAATATTTATTTTAGTAATAATG
TTTCAAAC TCATATATAACAATGCTGTATTTTAAGAATGATTACATAATGACTTATATTTGTATA
AAATAATTTTTATATTTGAAATGTTAACTTTTTATAGCACTAGCTATTTTAAAACAGGGGAGTGA
GGAGGACAGGGATGATAAGGATCATTCAACTTCATGTTGTGAAGACGAGCTGATGTAAATCTTGT
ACCCATCTGTGTGGTTCTCAGACAACACATGCTCTCTTTTAATGCAGCTTTGAAGAAGATGGTAC
CAAAGGTTAAGACGGCCCCCTGATGGGCACATCAACTTCTGAACTGCAAAC TAAGCTTTAGAGGA
ATGTATTATATTTATTACTGTAATAGAATATCATGTGTCAATAAAATCCTTTTATTTGTGTGAAA
>gi|90421312|ref|NM_000492.3| Homo sapiens cystic fibrosis
transmembrane conductance regulator (ATP-binding cassette
sub-family C, member 7) (CFTR), mRNA
AATTGGAAGCAAATGACATCACAGCAGGTCAGAGAAAAAGGGTTGAGCGGCAGGCACCCAGAGTA
GTAGGTCTTTGGCATTAGGAGCTTGAGCCCAGACGGCCCTAGCAGGGACCCCAGCGCCCGAGAGA
CCATGCAGAGGTCGCCTCTGGAAGAGGCCAGCGTTGTCTCCAACTTTTTTTCAGCTGGACCAGA
CCAATTTTGAGGAAAGGATACAGACAGCGCCTGGAATTGTCAGACATATACCAAATCCCTTCTGT
TGATTCTGCTGACAATCTATCTGAAAAATTGGAAAGAGAATGGGATAGAGAGCTGGCTTCAAAGA
AAAATCCTAACTCATTAATGCCCTTCGGCGATGTTTTTCTGGAGATTTATGTTCTATGGAATC
TTTTTATATTTAGGGGAAGTCACCAAAGCAGTACAGCCTCTCTTACTGGGAAGAATCATAGCTTC
CTATGACCCGGATAACAAGGAGGAACGCTCTATCGCGATTTATCTAGGCATAGGCTTATGCCTTC
TCTTTATTGTGAGGACACTGCTCCTACACCCAGCCATTTTTGGCCTTCATCACATTGGAATGCAG
ATGAGAATAGCTATGTTTAGTTTGATTTATAAGAAGACTTTAAAGCTGTCAAGCCGTGTTCTAGA
TAAAATAAGTATTGGACAACCTGTTAGTCTCCTTTCCAACAACCTGAACAAATTTGATGAAGGAC
TTGCATTGGCACATTTTCGTGTGGATCGCTCCTTTGCAAGTGGCACTCCTCATGGGGCTAATCTGG
GAGTTGTTACAGGCGTCTGCCTTCTGTGGACTTGGTTTCCTGATAGTCCTTGCCCTTTTTTCAGGC
TGGGCTAGGGAGAATGATGATGAAGTACAGAGATCAGAGAGCTGGGAAGATCAGTGAAAGACTTG
TGATTACCTCAGAAATGATTGAAAATATCCAATCTGTTAAGGCATACTGCTGGGAAGAAGCAATG
GAAAAAATGATTGAAAACCTTAAGACAAACAGAACTGAACTGACTCGGAAGGCAGCCTATGTGAG
ATACTTCAATAGCTCAGCCTTCTTCTCTCAGGGTTCTTTGTGGTGTTTTTATCTGTGCTTCCCT
ATGCACTAATCAAAGGAATCATCCTCCGGAAAATATTCACCACCATCTCATTCTGCATTGTTCTG
CGCATGGCGGTCACTCGGCAATTTCCCTGGGCTGTACAAACATGGTATGACTCTCTTGGAGCAAT
AAACAAAATACAGGATTTCTTACAAAAGCAAGAATATAAGACATTGGAATATAACTTAACGACTA
CAGAAGTAGTGATGGAGAATGTAACAGCCTTCTGGGAGGAGGGATTTGGGGAATTATTTGAGAAA
GCAAAACAAAACAATAACAATAGAAAACTTCTAATGGTGATGACAGCCTCTTCTTCAGTAATTT
CTCACTTCTTGGTACTCCTGTCCTGAAAGATATTAATTTCAAGATAGAAAGAGGACAGTTGTTGG
CGGTTGCTGGATCCACTGGAGCAGGCAAGACTTCACTTCTAATGGTGATTATGGGAGAACTGGAG
CCTTCAGAGGGTAAAATTAAGCACAGTGGAAGAATTTCAATTCTGTTCTCAGTTTTCTGGATTAT
GCCTGGCACCATTAAAGAAAATATCATCTTTGGTGTTTCCTATGATGAATATAGATACAGAAGCG
TCATCAAAGCATGCCAACTAGAAGAGGACATCTCCAAGTTTGCAGAGAAAGACAATATAGTTCTT
GGAGAAGGTGGAATCACACTGAGTGGAGGTCAACGAGCAAGAATTTCTTTAGCAAGAGCAGTATA
CAAAGATGCTGATTTGTATTTATTAGACTCTCCTTTTGGATACCTAGATGTTTTAACAGAAAAAG
AAATATTTGAAAGCTGTGTCTGTAACTGATGGCTAACAAAAC TAGGATTTTGGTCACTTCTAAA
ATGGAACATTTAAAGAAAGCTGACAAAATATTAATTTTGCATGAAGGTAGCAGCTATTTTTATGG
GACATTTTCAGAACTCCAAAATCTACAGCCAGACTTTAGCTCAAACTCATGGGATGTGATTCTT
TCGACCAATTTAGTGCAGAAAGAAGAAATTCATCCTAACTGAGACCTTACACCGTTTCTCATTA

GAAGGAGATGCTCCTGTCTCCTGGACAGAAACAAAAACAATCTTTTAAACAGACTGGAGAGTT
TGGGGAAAAAAGGAAGAATTCTATTCTCAATCCAATCAACTCTATACGAAAATTTTCCATTGTGC
AAAAGACTCCCTTACAAATGAATGGCATCGAAGAGGATTCTGATGAGCCTTTAGAGAGAAGGCTG
TCCTTAGTACCAGATTCTGAGCAGGGAGAGGCGATACTGCCTCGCATCAGCGTGATCAGCACTGG
CCCCACGCTTCAGGCACGAAGGAGGCAGTCTGTCCTGAACCTGATGACACACTCAGTTAACCAAG
GTCAGAACATTACCGAAAGACAACAGCATCCACACGAAAAGTGTCCTGGCCCCTCAGGCAAAC
TTGACTGAACTGGATATATATTCAAGAAGGTTATCTCAAGAACTGGCTTGGAATAAGTGAAGA
AATTAACGAAGAAGACTTAAAGGAGTGCTTTTTTGTGATGATATGGAGAGCATACCAGCAGTGACTA
CATGGAACACATACCTTCGATATATTACTGTCCACAAGAGCTTAATTTTTGTGCTAATTTGGTGC
TTAGTAATTTTTCTGGCAGAGGTGGCTGCTTCTTTGGTTGTGCTGTGGCTCCTTGGAACACTCC
TCTTCAAGACAAAGGGAATAGTACTCATAGTAGAAATAACAGCTATGCAGTGATTATCACCAGCA
CCAGTTCGTATTATGTGTTTTACATTTACGTGGGAGTAGCCGACACTTTGCTTGCTATGGGATTC
TTCAGAGGTCTACCACTGGTGCATACTCTAATCACAGTGTGCGAAAATTTTACACCACAAAATGTT
ACATTCTGTTCTTCAAGCACCTATGTCAACCCTCAACACGTTGAAAGCAGGTGGGATTCTTAATA
GATTCTCCAAAGATATAGCAATTTTGGATGACCTTCTGCCTCTTACCATATTTGACTTCATCCAG
TTGTTATTAATTGTGATTGGAGCTATAGCAGTTGTGCGAGTTTTACAACCCTACATCTTTGTTGC
AACAGTGCCAGTGATAGTGGCTTTTATTATGTTGAGAGCATATTTCTCCTCCAAACCTCACAGCAAC
TCAAACAACCTGGAATCTGAAGGCAGGAGTCCAATTTTCACTCATCTTGTTACAAGCTTAAAAGGA
CTATGGACACTTCGTGCCTTCGGACGGCAGCCTTACTTTGAACTCTGTTCCACAAAGCTCTGAA
TTTACATACTGCCAACTGGTTCTTGTACCTGTCAACACTGCGCTGGTTCCAAATGAGAATAGAAA
TGATTTTTGTGTCATCTTCTTCATTGCTGTTACCTTCATTTCCATTTTAACAACAGGAGAAGGAGAA
GGAAGAGTTGGTATTATCCTGACTTTAGCCATGAATATCATGAGTACATTGCAGTGGGCTGTAAA
CTCCAGCATAGATGTGGATAGCTTGATGCGATCTGTGAGCCGAGTCTTTAAGTTCATTGACATGC
CAACAGAAGGTAAACCTACCAAGTCAACCAAACCATACAAGAATGGCCAACTCTCGAAAGTTATG
ATTATTGAGAATTCACACGTGAAGAAAGATGACATCTGGCCCTCAGGGGGCCAAATGACTGTCAA
AGATCTCACAGCAAAATACACAGAAGGTGGAAATGCCATATTAGAGAACATTTCTTCTCAATAA
GTCCTGGCCAGAGGGTGGGCCTCTTGGGAAGAACTGGATCAGGGAAGAGTACTTTGTTATCAGCT
TTTTTGAGACTACTGAACACTGAAGGAGAAATCCAGATCGATGGTGTGTCTTGGGATTCAATAAC
TTTGCAACAGTGGAGGAAAGCCTTTGGAGTGATACCACAGAAAGTATTTATTTTTCTGGAACAT
TTAGAAAAAATTTGGATCCCTATGAACAGTGGAGTGATCAAGAAATATGGAAAGTTGCAGATGAG
GTTGGGCTCAGATCTGTGATAGAACAGTTTCCTGGGAAGCTTGACTTTGTCCTTGTGGATGGGGG
CTGTGTCCTAAGCCATGGCCACAAGCAGTTGATGTGCTTGGCTAGATCTGTTCTCAGTAAGGCGA
AGATCTTGCTGCTTGATGAACCCAGTGCTCATTTGGATCCAGTAACATACCAAATAATTAGAAGA
ACTCTAAAACAAGCATTTGCTGATTGCACAGTAATCTCTGTGAACACAGGATAGAAGCAATGCT
GGAATGCCAACAATTTTTGGTCATAGAAGAGAACAAGTGCGGCAGTACGATTCCATCCAGAAAC
TGCTGAACGAGAGGAGCCTCTTCCGGCAAGCCATCAGCCCCCTCCGACAGGGTGAAGCTCTTTCCC
CACCGGAACTCAAGCAAGTGCAAGTCTAAGCCCCAGATTGCTGCTCTGAAAGAGGAGACAGAAGA
AGAGGTGCAAGATACAAGGCTTTAGAGAGCAGCATAAATGTTGACATGGGACATTTGCTCATGGA
ATTGGAGCTCGTGGGACAGTCACCTCATGGAATTGGAGCTCGTGGAACAGTTACCTCTGCCTCAG
AAAACAAGGATGAATTAAGTTTTTTTTTAAAAAAGAAACATTTGGTAAGGGGAATTGAGGACACT
GATATGGGTCTTGATAAATGGCTTCCTGGCAATAGTCAAATTGTGTGAAAGGTAATTCAAATCCT
TGAAGATTTACCACTTGTTGTTTTGCAAGCCAGATTTTCTGAAAACCCTTGCCATGTGCTAGTAA
TTGGAAAGGCAGCTCTAAATGTCAATCAGCCTAGTTGATCAGCTTATTGTCTAGTGAACTCGTT
AATTTGTAGTGTTGGAGAAGAACTGAAATCATACTTCTTAGGGTTATGATTAAGTAATGATAACT
GGAAACTTCAGCGGTTTATATAAGCTTGTATTCCTTTTTCTCTCCTCTCCCCATGATGTTTAGAA
ACACAACATATATTGTTTGCTAAGCATTCCAACATCTCATTTCCAAGCAAGTATTAGAATACCAC
AGGAACCACAAGACTGCACATCAAAATATGCCCCATTCAACATCTAGTGAGCAGTCAGGAAAGAG


```

AACTTCCAGATCCTGGAAATCAGGGTTAGTATTGTCCAGGTCTACCAAAAATCTCAATATTTTCAG
ATAATCACAATACATCCCTTACCTGGGAAAGGGCTGTTATAATCTTTCACAGGGGACAGGATGGT
TCCCTTGATGAAGAAGTTGATATGCCTTTTCCCAACTCCAGAAAGTGACAAGCTCACAGACCTTT
GAACTAGAGTTTAGCTGGAAAAGTATGTTAGTGCAAATTGTACAGGACAGCCCTTCTTTCCACA
GAAGCTCCAGGTAGAGGGTGTGTAAGTAGATAGGCCATGGGCACTGTGGGTAGACACACATGAAG
TCCAAGCATTTAGATGTATAGGTTGATGGTGGTATGTTTTTCAGGCTAGATGTATGTACTTCATGC
TGTCTACACTAAGAGAGAATGAGAGACACACTGAAGAAGCACCAATCATGAATTAGTTTTATATG
CTTCTGTTTTATAATTTTGTGAAGCAAAATTTTTCTCTAGGAAATATTTATTTTAATAATGTTT
CAAACATATATAACAATGCTGTATTTTAAAAGAATGATTATGAATTACATTTGTATAAAATAATT
TTTATATTTGAAATATTGACTTTTTATGGCACTAGTATTTCTATGAAATATTATGTTAAACTGG
GACAGGGGAGAACCTAGGGTGATATTAACCAGGGGCCATGAATCACCTTTTGGTCTGGAGGGAAG
CCTTGGGGCTGATGCAGTTGTTGCCACAGCTGTATGATTCCCAGCCAGCACAGCCTCTTAGATG
CAGTTCTGAAGAAGATGGTACCACCAGTCTGACTGTTTCCATCAAGGGTACACTGCCTTCTCAAC
TCCAAACTGACTCTTAAGAAGACTGCATTATATTTATTACTGTAAGAAAATATCACTTGTCAATA
AAATCCATACATTTGTGTGAAA

```

Figura 9.21. Secuencias bajo estudio

Una vez introducidas las secuencias y, puesto que son relativamente largas, ajustaremos los parámetros del gráfico como los de la Figura 9.22 con un *zoom* de 1:10 y un tamaño de ventana de 59:

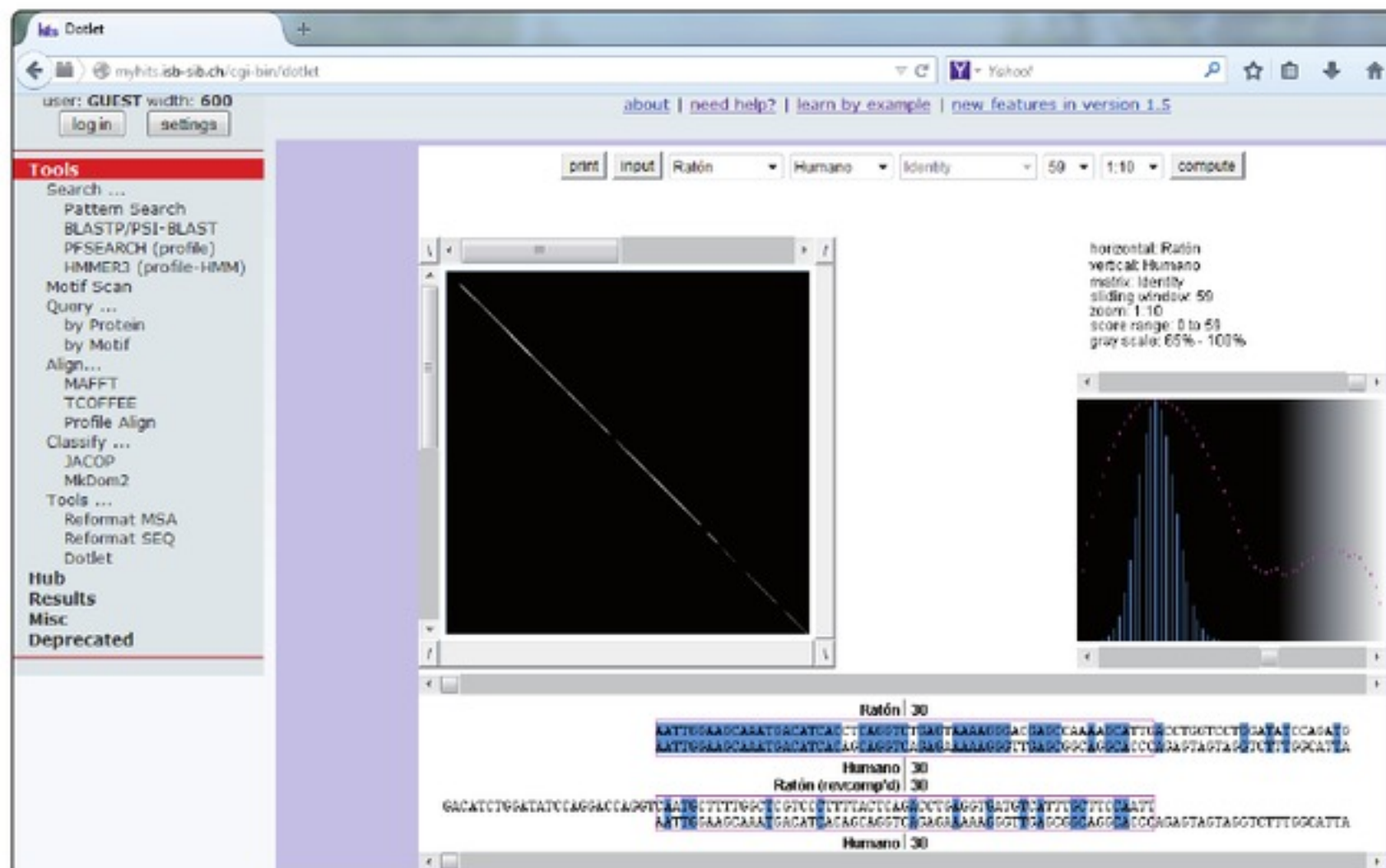


Figura 9.22 Gráfico de puntos de dos secuencias diferentes

9.2.4 Búsqueda de ORF



Una ORF (*Open Reading Frame*) es la secuencia de ADN comprendida entre un codón de inicio (generalmente, ATG) y un codón de parada (TAA, TAG o TGA).

9.2.4.1 NCBI-ORFFINDER

La herramienta básica más utilizada para predecir la presencia de secuencias codificantes en proteínas es el ORF Finder del NCBI. Se trata de una herramienta gráfica de análisis para la búsqueda de ORF de tamaño seleccionable en una secuencia, ya sea introducida por el usuario (y, potencialmente desconocida) o recuperada de una base de datos de secuencias.

En este primer ejemplo, utilizaremos la siguiente secuencia:

```
>Secuencial Libro Bioinformatica
GTTGCCCCGGCCGCGCCGCGCGCGCTACCTATCTCGCCGCGCGGCCTCGACCCCTGCTTC
CGTCTGGCTGCCTGTGCCGCGTGGTGCCGGACCCGGGGCAGTGTGCAGGGCCGCCGGGAAAGGGA
AGGAGGTGCTCAGCGGCGTGGTCTTCCAGCCATTCGAGGAG CTCAAGGGGGAGCTCTCCCTCGT
CCCCCAGGCCAAGGACCAGTCTCTCGCTAGGCAAAAGTTCGTGACGAGTGCGAGGCCGCCATCA
ACGAGCAGATCAATGTGGAGTACAATGCATCGTACGCGTACCACTCCCTTTTCGCCTACTTTGAT
CGTGACAACGTTGCTCTCAAGGGATTGCGCAAATTCTTCAAAGAATCCAGCGATGAGGAGAGGGA
TCACGCAGAGAACTCATCAAGTACCAGAACATGCGTGGAGGCAGGGTGCGGCTCCAGTCCATCG
TCACACCTTTGACAGAGTTGACCATCCTGAGAAAGGGGATGCCTTGTATGCTATGGAGTTGGCC
TTGGCTCTCGAAAAGCTTGTAATGAGAAGTTGCACAACCTGCACAGTGTGGCATCAAGGTGCAA
TGATCCACAGCTGACCGACTTCGTTGAGAGCGAATTCCTTGAGGAGCAGGTTGAAGCCATCAAGA
AGATCTCTGAGTATGTCGCCAGCTGAGAAGAGTGGGAAAGGGGCATGGGGTGTGGCACTTTGAT
CAGAAGCTGCTTGAGGAAGAAGCT
```

Figura 9.25. Secuencia bajo estudio

Abrimos un navegador e introducimos la página principal del ORF Finder: <http://www.ncbi.nlm.nih.gov/gorf/gorf.html>. Una vez allí, pegue la secuencia anterior en el área de texto destinada para ello (ver Figura 9.26) o, si se trata de una secuencia obtenida de alguna base de datos (GenBank, por ejemplo), se puede introducir el GI o el *accession*.

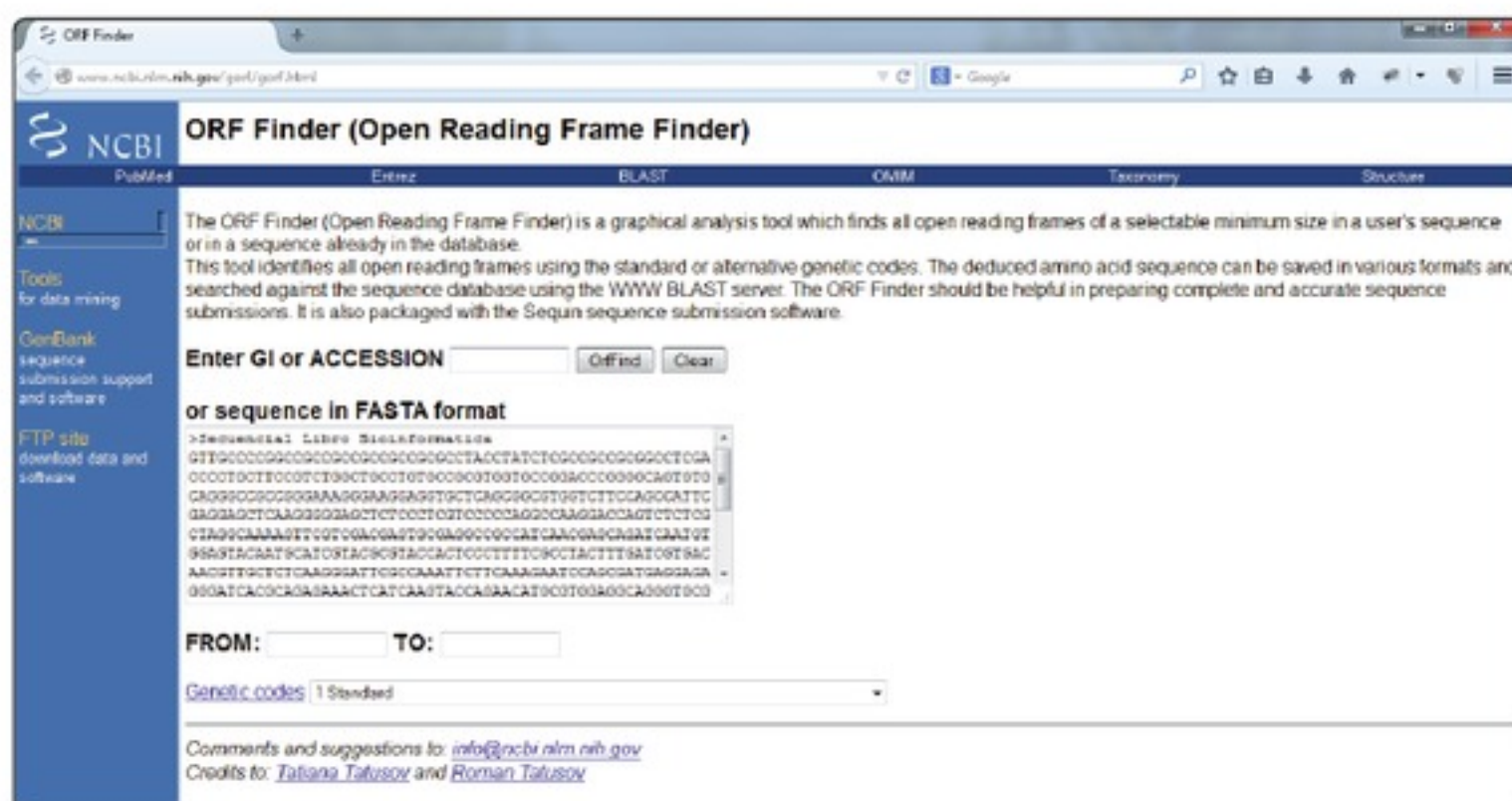


Figura 9.26. Página principal del ORF Finder

Junto al enlace **Genetic codes** hay un menú desplegable que permite seleccionar el diccionario genético de codones. Por defecto, se emplea el estándar aunque en el menú se dispone de hasta 22 opciones.



En el enlace hay una descripción detallada del diccionario de traducción.

<http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c>

Pulsamos el botón **OrfFind** y aparecerá la ventana de resultados (ver Figura 9.27).

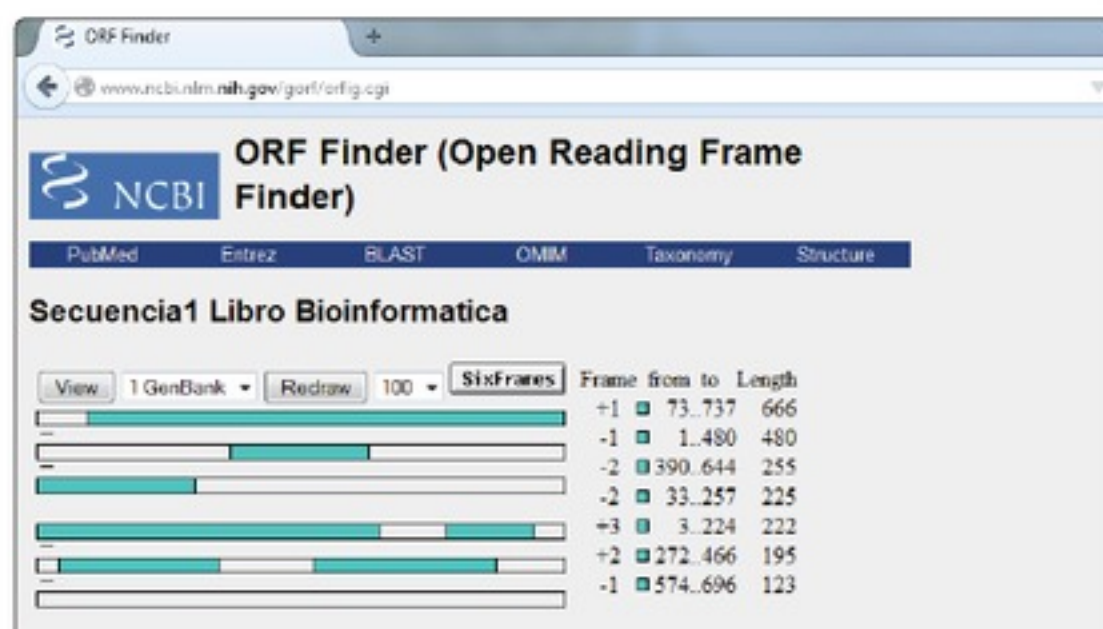


Figura 9.27. Ventana de resultados del ORF Finder

Si se hace clic en una ORF, su zona en la imagen cambiará a violeta, indicando que se ha seleccionado, y aparecerán las secuencias de bases y de aminoácidos. Por ejemplo, si pinchamos en la primera ORF, obtendremos la imagen de la Figura 9.28.



Si ahora se pincha en el botón **Accept**, la región seleccionada se vuelve verde claro indicando que hemos aceptado la ORF y que estamos interesados en obtener más información acerca de la misma.

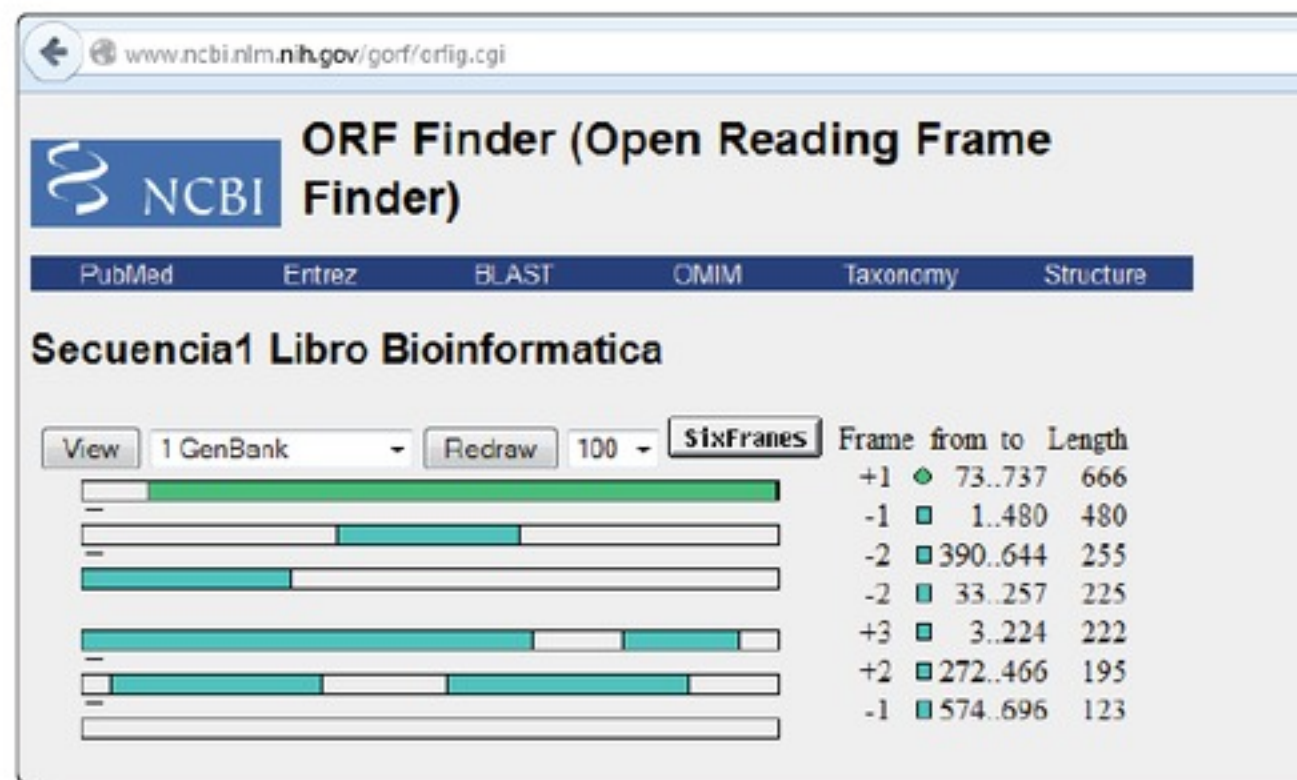


Figura 9.29. Información sobre la región seleccionada

Por ejemplo, podríamos ver la información de la ORF en formato GenBank pulsando el botón **View** con la opción de GenBank seleccionada:



Figura 9.30. Registro en GenBank

9.2.4.2 GENEMARK

Los algoritmos de búsqueda de ORF convencionales identifican alrededor del 85% de las regiones codificantes de interés. Sin embargo, existen situaciones en las que es necesario una aproximación más compleja. Tal es el caso de la búsqueda de proteínas cortas, la resolución de ambigüedades o la identificación del codón de inicio exacto. Una de las herramientas que utilizan estos algoritmos más complejos es GeneMark (<http://exon.gatech.edu/GeneMark/>).

Aprenderemos a utilizar GeneMark con la secuencia del gen UTPase de la *Escherichia coli*, cuyo *accession* es el X01714. Podemos obtener su secuencia directamente desde GenBank:

```
>gi|41296|emb|X01714.1| E. coli dut gene for dUTPase (EC 3.6.1.23)
(deoxyuridine 5'-triphosphate nucleotidohydrolase)
CAGAGAAAATCAAAAAGCAGGCCACGCAGGGTGATGAATTAACAATAAAAATGGTTAAAAACCCC
GATATCGTCGCAGGCGTTGCCGCACTAAAAGACCATCGACCCTACGTCGTTGGATTTGCCGCCGA
AACAAATAATGTGGAAGAATACGCCCGGCAAAAACGTATCCGTAAAAACCTTGATCTGATCTGCG
CGAACGATGTTTCCCAGCCAACTCAAGGATTTAACAGCGACAACAACGCATTACACCTTTTCTGG
CAGGACGGAGATAAAGTCTTACCGCTTGAGCGCAAAGAGCTCCTTGGCCAATTATTACTCGACGA
GATCGTGACCCGTTATGATGAAAAAATCGACGTTAAGATTCTGGACCCGCGCGTTGGGAAGGAA
TTTCCGCTCCCGACTTATGCCACCTCTGGCTCTGCCGGAAGTTGACCTGCGTGCCTGTCTCAACGA
CGCCGTAGAACTGGCTCCGGGTGACACTACGCTGGTTCGACCGGGCTGGCGATTATATTGCCG
ATCCTTCACTGGCGGCAATGATGCTGCCGCGCTCCGGATTGGGACATAAGCACGGTATCGTGCTT
GGTAACCTGGTAGGATTGATCGATTCTGACTATCAGGGCCAGTTGATGATTTCCGTGTGGAACCG
TGGTCAGGACAGCTTCACCATTCACCTGGCGAACGCATCGCCCAGATGATTTTTTGTTCGGTAG
TACAGGCTGAATTTAATCTGGTGGAAGATTTTCGACGCCACCGACCGCGGTGAAGGCGGCTTTGGT
CACTCTGGTCGTCAGTAACACATACGCATCCGAATAACGTCATAACATAGCCGCAAACATTTTCGT
TTGCGGTCATAGCGTGGGTGCCGCTGGCAAGTGCTTATTTTCAGGGGTATTTTGTAACATGGCA
GAAAAACAACTGCGAAAAGGAACCGTCGCGAGGAAATACTTCAGTCTCTGGCGCTGATGCTGGA
ATCCAGCGATGGAAGCCAACGTATCACGACGGCAAACTGGCCGCCTCTGTGCGGCTTTCCGAAG
CGGCACTGTATCGCCACTTCCCCAGTAAGACCCGCATGTTTCGATAGCCTGATTGAGTTTATCGAA
GATAGCCTGATTACTCGCATCAACCTGATTCTGAAAGATGAGAAAGACACCACAGCGCGCCTGCG
TCTGATTGTGTTGCTGCTTCTCGGTTTTGGTGAGCGTAATCCTGGCCTGACCCGCATCCTCACTG
GTCATGCGCTAATGTTTGAACAGGATCGCCTGCAAGGGCGCATCAACCAGCTGTTTCGAGCGTATT
GAAGCGCAGCTGCGCCAGGTATTGCGTGAAAAGAGAATGCGTGAGGGTGAAGGTTACACCACCGA
TGAAACCCTGCTGGCAAGCCAGATCCTGGCCTTCTGTGAAGGTATGCTGTACGTTTTGTCCGCA
GCGAATTTAAATACCGCCCGACGGATGATTTTGACGCCCGCTGGCCGCTAATTGCGGGCAGTTGC
AGTAATATGACGCCGGATGACTTTTCATCCGGCGAGTTTCTTTAAACGCCAACTCTTCGCGATA
GGCCTTAACCGCCGCCAGATGTTCCGCCATTTCCGGCTTCTCTTCCAGG
```

Figura 9.31. Secuencia bajo estudio

El primer paso será abrir la página principal de GeneMark (ver Figura 9.32) y seleccionar el tipo de análisis que se desea hacer. Existen varias versiones especializadas del programa, cada una de las cuales se corresponde con un modelo distinto de gen.

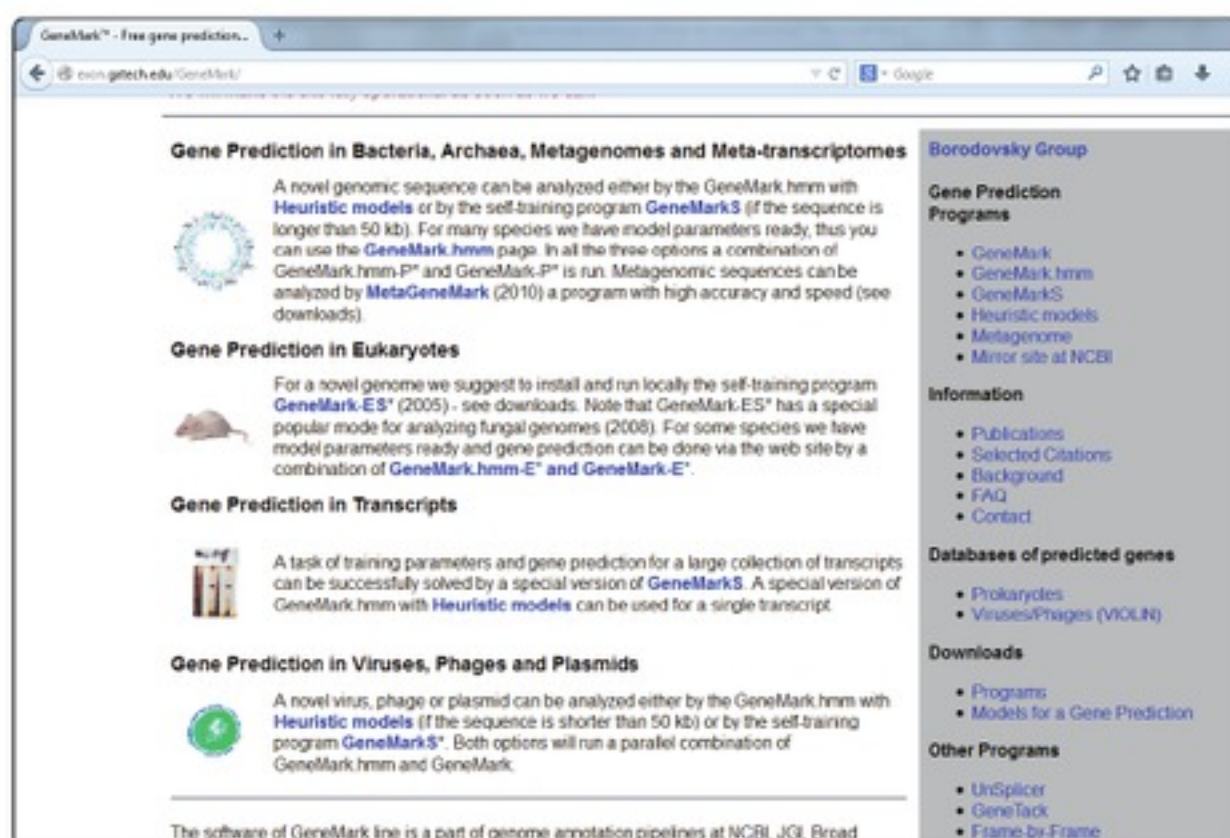


Figura 9.32. Página principal de GeneMark

Supondremos que se desea predecir un gen en una bacteria, por lo que se seleccionará el enlace **Gene Prediction in Bacteria, Archaea and Meta-transcriptomes** y pinchamos en el enlace **GeneMarkS** (ver Figura 9.33). Además de la secuencia podemos especificar algunas opciones de búsqueda como el tipo de secuencia (en nuestro caso, procariota) o el formato de salida. Seleccionaremos todas las opciones de salida (secuencia de proteínas, secuencia de nucleótidos de los genes y PDF).

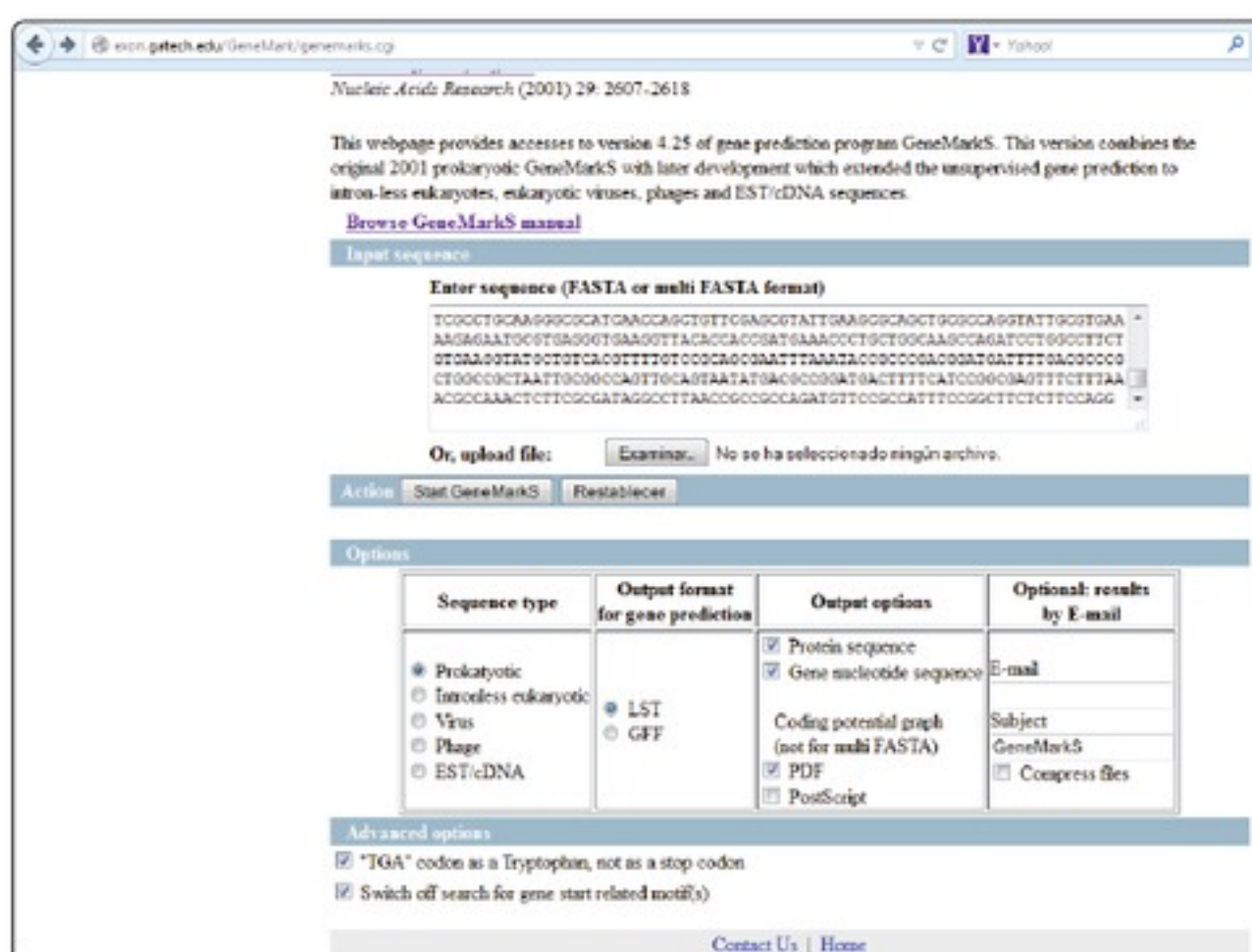


Figura 9.33. Página principal de GeneMarkS

Si pinchamos en el botón **Start GeneMarkS** obtenemos una página de resultados como la de la Figura 9.34:

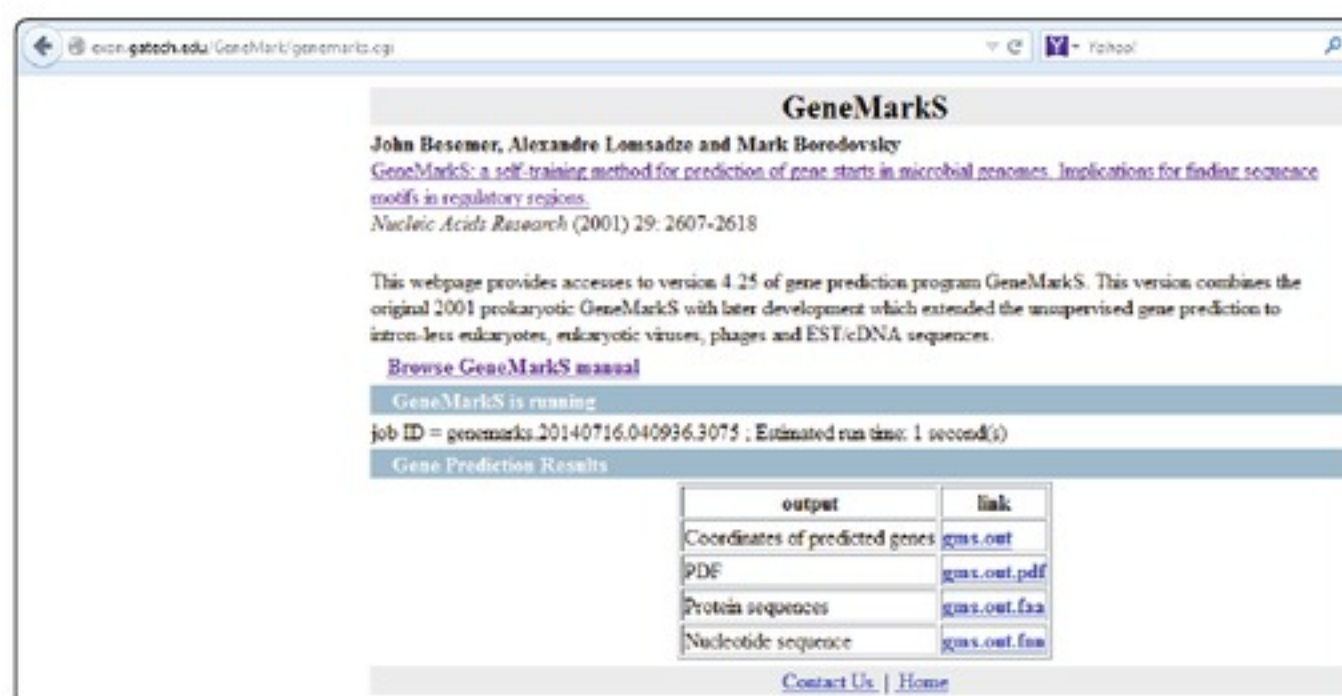


Figura 9.34. Página de resultados

En primer lugar, examinaremos la sección de predicción de genes navegando por el gms.out (ver Figura 9.35). En la cabecera del informe de resultados se encuentran datos básicos del mismo entre los que se encuentran la fecha, el algoritmo utilizado para hacer la predicción, etc. Por otra parte, bajo la definición de la secuencia que hemos introducido, se encuentran las coordenadas de los genes, indicando la hebra, las posiciones de inicio y final.



Figura 9.35. Descripción de los resultados

Si comparamos la predicción con la información contenida en el registro de GenBank, veremos que en esta base de datos se indica que esta secuencia contiene dos regiones CDS, una alrededor de la posición 340 y otra alrededor de la 900.

```
LOCUS      X01714                1609 bp  DNA   linear  BCT 23-OCT-2008
DEFINITION E. coli dut gene for dUTPase (EC 3.6.1.23)
(deoxyuridine
          5'-triphosphate nucleotidohydrolase).
ACCESSION  X01714
```



```

VERSION    X01714.1 GI:41296
KEYWORDS   dUTPase; unidentified reading frame.
SOURCE     Escherichia coli
ORGANISM   Escherichia coli
            Bacteria; Proteobacteria; Gammaproteobacteria;
            Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE  1 (bases 1 to 1609)
AUTHORS    Lundberg,L.G., Thoresson,H.O., Karlstrom,O.H. and
            Nyman,P.O.
TITLE      Nucleotide sequence of the structural gene for dUTPase
            of Escherichia coli K-12
JOURNAL    EMBO J. 2 (6), 967-971 (1983)
PUBMED     6139280
COMMENT    Data kindly reviewed (25-NOV-1985) by L. Lundberg.
FEATURES   Location/Qualifiers
            source          1..1609
                        /organism="Escherichia coli"
                        /mol_type="genomic DNA"
                        /db_xref="taxon:562"
            promoter        286..291
                        /note="-35 region"
            promoter        310..316
                        /note="-10 region"
            misc_feature     322..324
                        /note="put. transcription start region"
            RBS              330..333
                        /note="put. rRNA binding site"
            CDS              343..798
[...]
```

```

            CDS              905..1540
[...]
```

Figura 9.36. Registro GenBank

En los resultados sobre las secuencias de proteínas (enlace gms.out.faa), se muestra la secuencia de aminoácidos correspondiente a cada uno de los genes que se han predicho en formato FASTA (ver Figura 9.37):

```

>gene_1[GeneMark.hmm|119_aa|+|31362] >gi|41296|emb|X01714.1| E. coli dut gene for dUTPase (EC 3.6.1.23) (deoxyuridine 5'-triphosphate nucleotidohydrolase)
EKIKKQATQSGELTIMKVKHFDIVAGVAALKDERFYVSTAACTHNVVEYARQKIKKML
DLICANDVSGPTQSPNJOHNLHLEWQGGHVLPLERELLOQLLDEIVTRYDERRR
>gene_2[GeneMark.hmm|152_aa|+|3401798] >gi|41296|emb|X01714.1| E. coli dut gene for dUTPase (EC 3.6.1.23) (deoxyuridine 5'-triphosphate nucleotidohydrolase)
MKKKIDVKKLDFPVKKEFFLPTATSGGAGLDLRACLNDAVELAPGDTLVPFLAINIA
DPSLARDMLPSSGLCKHKGIVLGNLUGLIDSDVQQLMISVWNRQDSFTIQDGERIAQM
IIVPVVQAEFNLVIDFEDATGEGCGFNGSCFQ
>gene_3[GeneMark.hmm|211_aa|+|9051540] >gi|41296|emb|X01714.1| E. coli dut gene for dUTPase (EC 3.6.1.23) (deoxyuridine 5'-triphosphate nucleotidohydrolase)
MAERQTAKRGRGKEITLQSLAMESSDGSQITTAALAASVGVSKAALYHMFPSKTHMFD
SLIEFIEDSLITRINLILMDKDTIARLRILVLLIGFGRHSGULTRILTGMALNFEQGR
LQQRINGLPERIEAQLRQVLAERHNRGEGDYTTDETLLASQILAFCEHLSRFVKEFPY
RPTDDFDARMTLIAASCNRITPDDFSSSEFL
```

Figura 9.37. Secuencias de proteínas de los genes

Figura 9.39. Gráficas con los genes predichos

De esta gráfica, podemos obtener información muy interesante, tal y como muestra la Figura 9.40:

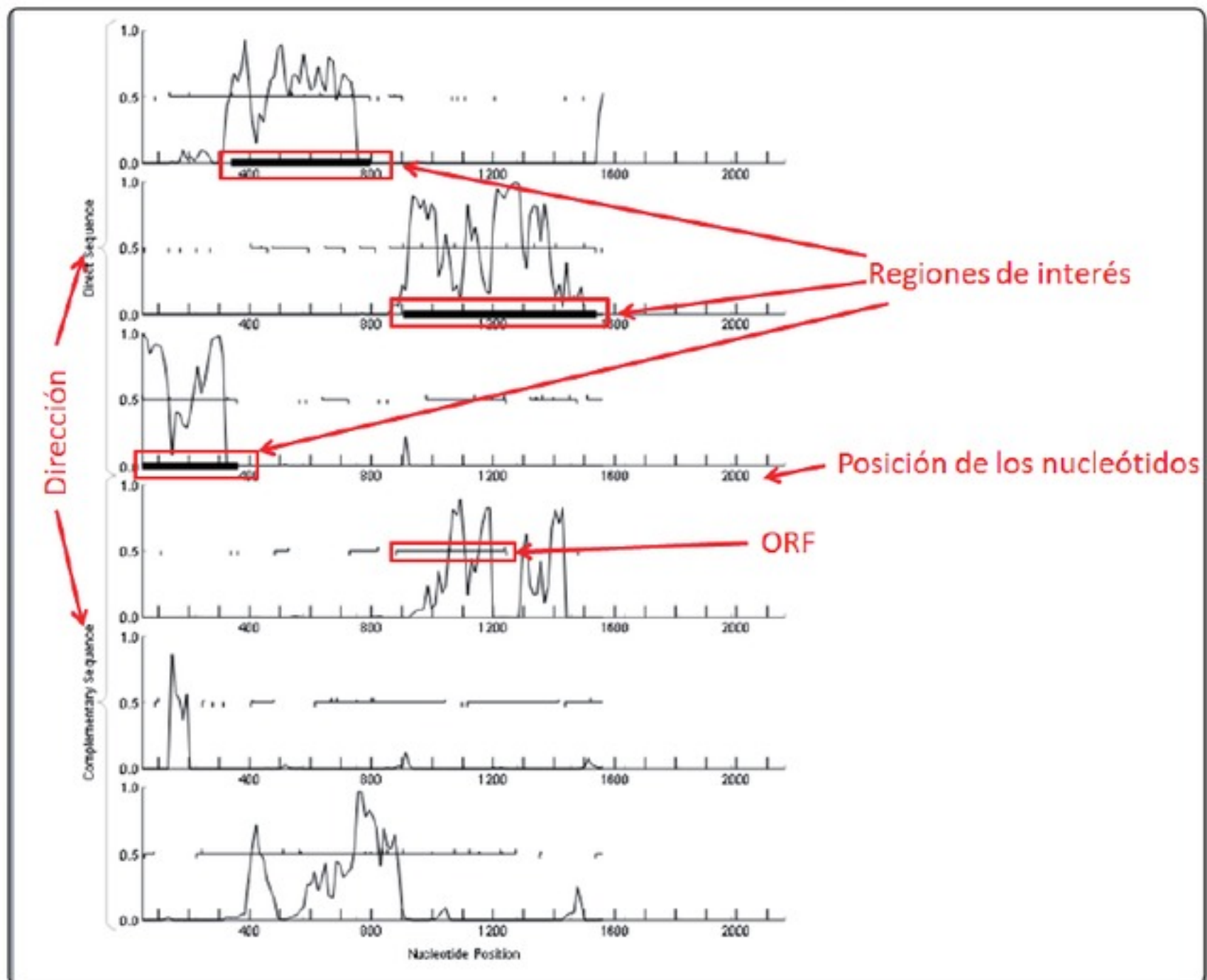


Figura 9.40. Análisis de la gráfica



Como acabamos de ver, con GeneMark es posible encontrar ORF. Compara los resultados con los obtenidos con el NCBI-ORF Finder.

9.3 ALINEAMIENTO DE SECUENCIAS CON BLASTN

Dos secuencias homólogas, es decir, con un ancestro común, suelen tener la misma estructura 3D y funciones relacionadas. La mejor manera de encontrar una secuencia homóloga a la secuencia bajo estudio es buscarla en una base de datos a través de BLAST. Una vez encontrada, para decidir si son homólogas o no, conviene tener en cuenta los siguientes dos criterios:

Secuencia de ADN: al menos el 70% de similitud en más de 100 bases entre la referencia y la secuencia bajo estudio o un valor E menor de 10^{-4} .

- Secuencia proteica: más del 25% de similitud en más de 100 aminoácidos entre la referencia y la secuencia bajo estudio o un valor E menor de 10^{-4} .
- Sin embargo, estos umbrales son orientativos: si una referencia se encuentra cerca del umbral, podría ser homóloga o no serlo. Es aquí cuando las técnicas de alineamiento de pares ayudan a decidir la relevancia de la referencia.

BLASTN se emplea para comparar una secuencia de nucleótidos bajo estudio con la información guardada en una base de datos de secuencias de nucleótidos. La situación más habitual es aquella en la que llega al laboratorio una secuencia (la secuencia bajo estudio) y se desea identificarla, comparándola con otras secuencias conocidas.

Se trata de identificar la siguiente secuencia:

```
>Anonimal
CAGGCAGCCCCACACCCTCCGCCTCCTGCACCGAGAGACATGGAATAAAGCCCCTGAACCAGCCC
TGCTGTGCGTCTGTGTGTCTTGGGGGCCCTGGGCCAAGCCCCACTTCCCGGCACTGTTGTGAGC
CCCTCCCAGCTCTCTCCATGCTCTCTGGGTGCCACAGGTGCCAACGCCAGCCAGGCCAGCCAGCATG
CAGTGGCTCTCCCCAAAGCGGCCATGCCTGTGCGCTGCCTGCTACCCCCACCCTGTGGCTCAGGG
TCCAGTATGGGAGCTGCGGGGGTCTCTGAGGGGCCAGGGGTGGTGGGGCCACTGAGAAATGACTT
CTTGTTTCAGTAGCTCTGGACTCTTGGAGTCCCCAGAGACCTTGTTTCAGGAAAGGGAATGAGAACA
TTCCAGCAATTTTCCCCCACCTAGCCCTCCCAGGTTCTATTTTATGATTTATTTCTGATGGAGT
CCCTGTGGAGGGAGGAGGCTGGGCTGAGGGAGGGGGTCTGCAGGGCGGGGGGCTGGGAAGGTGG
GGAGAGGCTGCTGAGAGCCACCCGCTATCCCCAGCTCTGGGCAGCCCTGGGACAGTCACACACCC
TGGCCTCGCGGCCCAAGCTGGCAGCCGTCTGCAGCCACAGCTTATGCCAGCCCAGGTCCAGCCAG
ACACCTGAGGGACCCACTGGTGCCTTGGAGGAAGCAGGAGAGGTCAGATGGCACCATGAGCTGGG
GCAGGTGCAGGGACCGTGGCAGCACCGGG
```

Figura 9.41. Secuencia bajo estudio

En primer lugar, abrimos un navegador e introducimos la página principal de NCBI-BLAST: <http://blast.ncbi.nlm.nih.gov/Blast.cgi> y aparecerá la pantalla de la Figura 9.42:

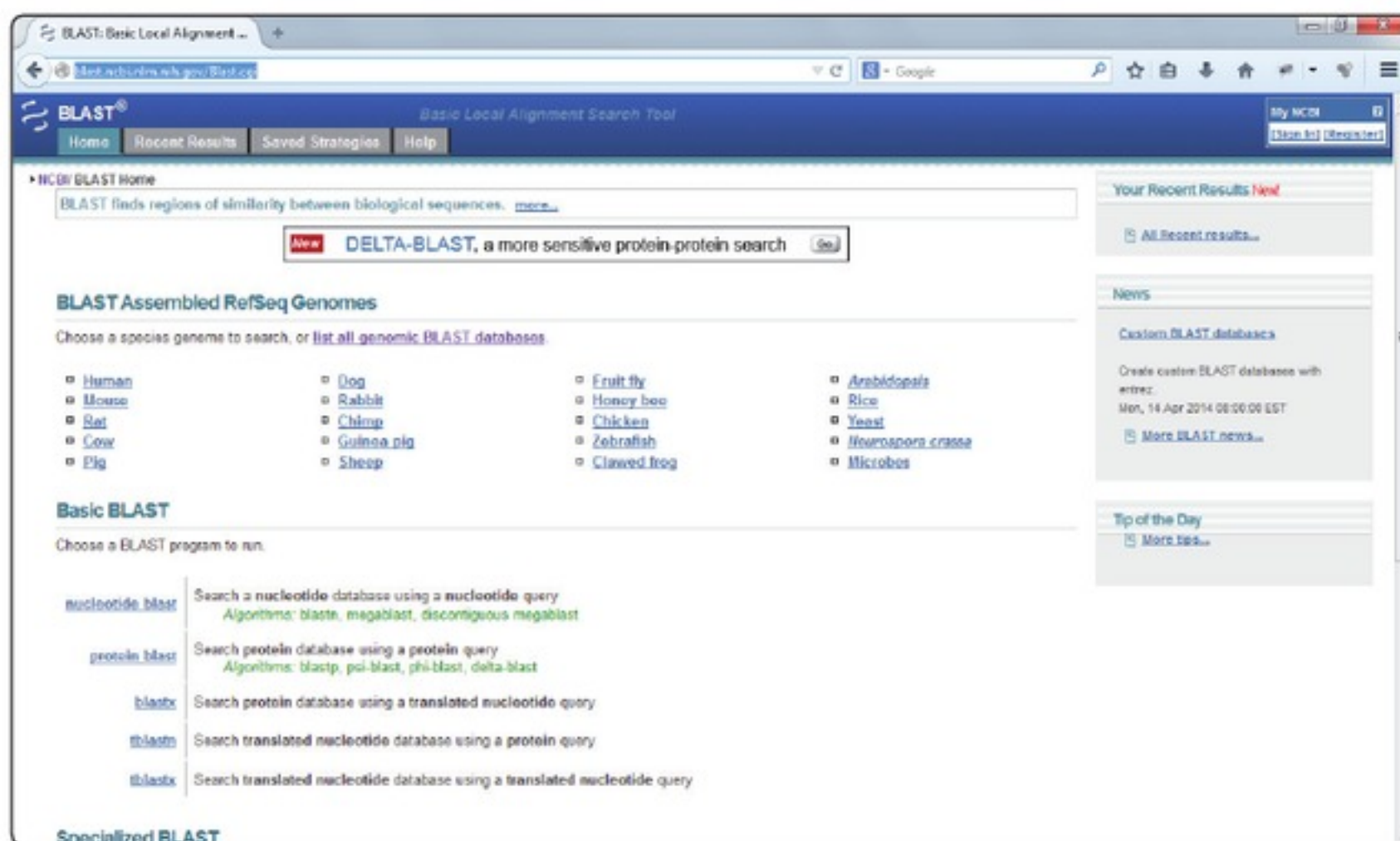


Figura 9.42. Página principal de BLAST

Seleccionamos BLASTN pinchando en el enlace **nucleotide blast** y aparecerá la ventana de introducción de los parámetros búsqueda. Una vez allí, el paso siguiente es especificar los datos de la secuencia bajo estudio. Aunque aquí pegaremos la secuencia en el área de texto, en general, existen tres métodos para indicar la secuencia de análisis:

- Escribir el identificador GI, si la secuencia ya se encuentra en una base de datos.
- Si la secuencia no está registrada en una base de datos (o se trata de una secuencia en bruto), podemos subir el fichero FASTA de la misma o pegarla en el área de texto, tal y como se hará en esta práctica.
- Introducir el *accession* de la secuencia.

Standard Nucleotide BLAST

Enter Query Sequence

Enter accession number(s), g(s), or FASTA sequence(s)

Query subrange

From

To

Or, upload file

Examine...

No se ha seleccionado ningún archivo

Job Title

Anonymous

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database

Human genomic + transcript

Mouse genomic + transcript

Others (nr etc.)

Nucleotide collection (nr/nr)

Organism

Optional

Enter organism name or id - completions will be suggested

Exclude

Optional

Models (3M7P)

Uncultured/environmental sample sequences

Entrez Query

Optional

Enter an Entrez query to limit search

Program Selection

Optimize for

Highly similar sequences (megablast)

More dissimilar sequences (discontiguous megablast)

Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST

Search database Nucleotide collection (nr/nr) using Megablast (Optimize for highly similar sequences)

Show results in a new window

Figura 9.43. Identificación de secuencias con BLASTN

Además de la secuencia o un rango de la misma, también es posible especificar otros parámetros. En la sección **Choose Search Set**, se puede restringir la búsqueda a un conjunto de datos concreto, como el genoma humano o EST. Por otra parte, la sección **Program Selection** permite escoger el algoritmo BLAST concreto que se empleará para calcular el alineamiento entre tres opciones:

- Megablast: está optimizado para secuencias con un alto grado de similitud (el 95% o más). Es la opción por defecto.
- Megablast discontinuo: utiliza una semilla inicial que ignora algunas bases (permite no coincidencias) y está pensado para comparaciones de secuencias pertenecientes a especies distintas.
- BLASTN: es el más lento y el que maneja grados de similitud más bajos.

Dejaremos la opción por defecto (Megablast) y al final de la práctica repetiremos el análisis para observar los resultados. Completados todos los parámetros de búsqueda, pulsamos el botón **BLASTN** para obtener los resultados.

La Figura 9.44 muestra los resultados que obtiene BLASTN y que se dividen en tres grandes partes: cabecera, gráficos, lista de coincidencias y alineamientos.

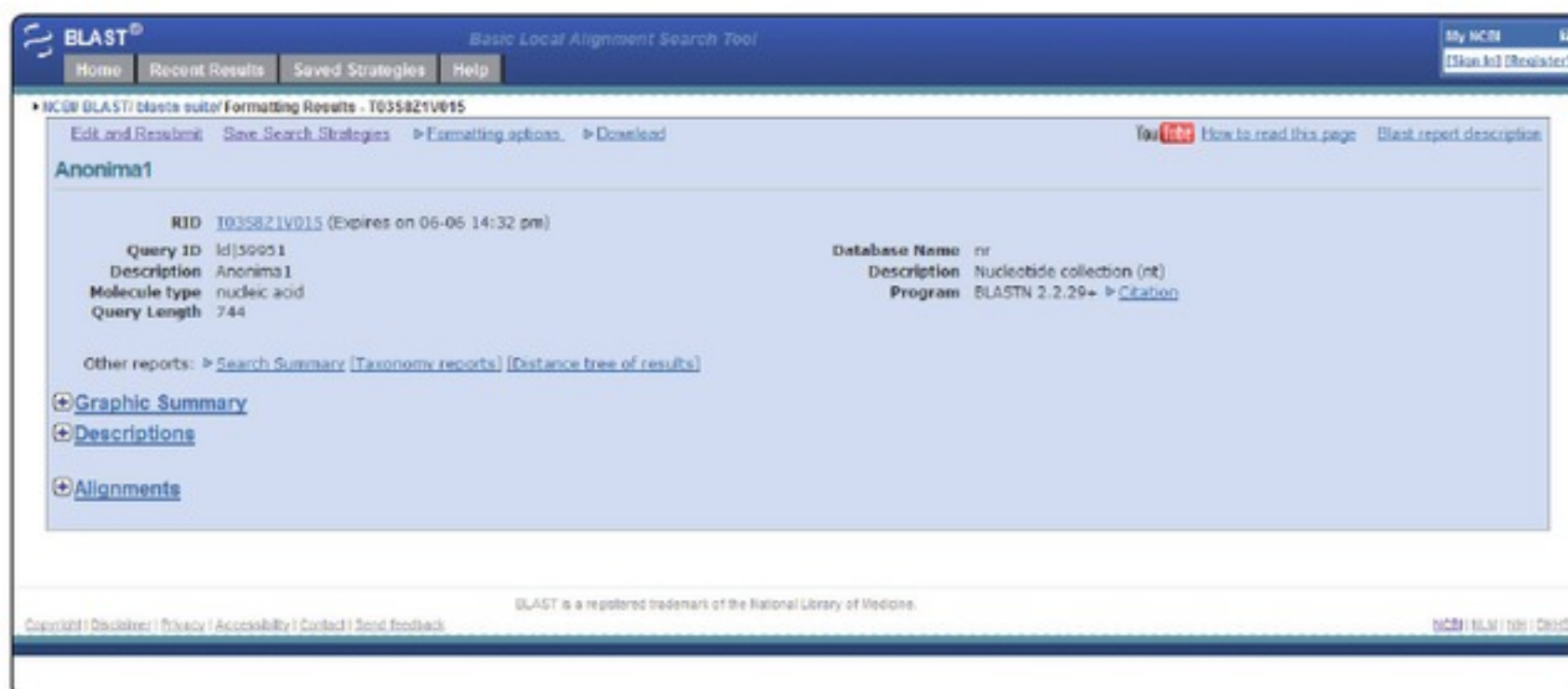


Figura 9.44. Informe de resultados de BLASTN

En la cabecera del informe de salida (ver Figura 9.45) encontramos información específica del programa (versión, compilación, fecha, etc.) y una lista de referencias científicas que deben incluirse si se publica alguna investigación que utilice BLAST. Sin embargo, la información de la cabecera más importante para la identificación de la secuencia son los nombres de la secuencia y el de la base de datos.



Figura 9.45. Cabecera BLASTN

La segunda parte de la página de resultados proporciona una vista gráfica (ver Figura 9.46) de las coincidencias encontradas en la base de datos y del grado de similitud con cada una de ellas. La secuencia bajo estudio se representa como un

conjunto de líneas gruesas en la parte superior de la figura, justo bajo la clave del código de colores, mientras que las coincidencias encontradas en la base de datos aparecen justo debajo, de tal manera que la más parecida ocupa la posición superior.

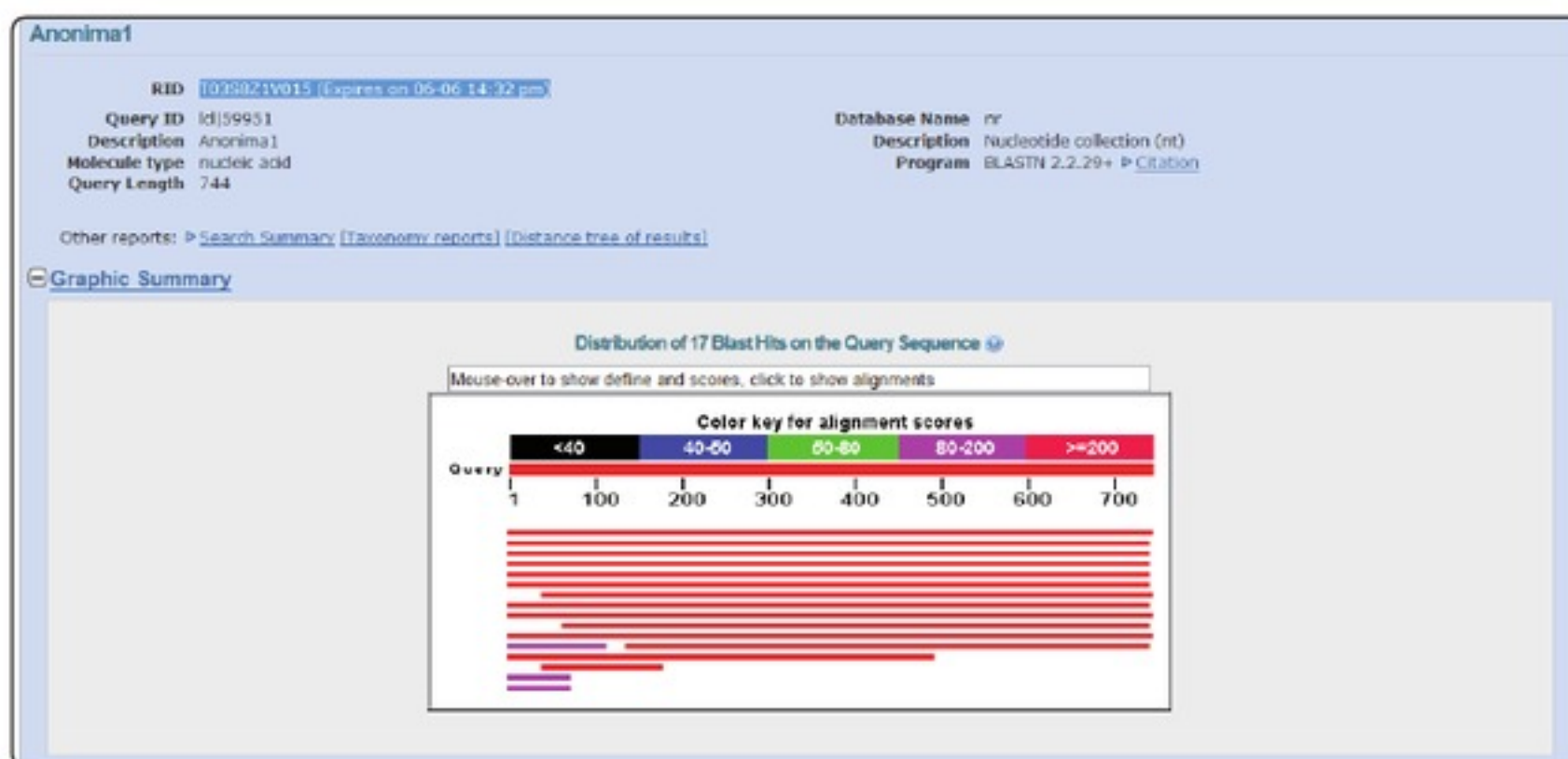


Figura 9.46. Resumen gráfico de las coincidencias con la base de datos

Si se mueve el ratón por la gráfica, se va mostrando la secuencia con la que existe coincidencia en ese fragmento de la secuencia bajo estudio. La interpretación de la gráfica es sencilla. El color rojo significa un alto grado de coincidencia entre la secuencia registrada en la base de datos y la secuencia bajo estudio; si el color es rosa, la coincidencia es moderadamente alta, mientras que un color verde indica que es una coincidencia poco significativa. Los valores más bajos son el azul y el negro. Finalmente, si se pincha en la barra del gráfico, BLASTN muestra el alineamiento correspondiente a la secuencia asociada al área en que se hizo clic.

La siguiente sección del informe de resultados es la que contiene la información detallada de las coincidencias con las secuencias registradas en la base de datos de nucleótidos o *hit list* (ver Figura 9.47). BLAST ordena las coincidencias calculando una medida estadística de similitud llamada *E-value* (valor esperado). Cuanto más bajo es el valor de E, menos significativa es la coincidencia. De la misma manera, un alto valor de E indica que las secuencias han evolucionado de un ancestro común (secuencias homólogas).

Order reports: [Position Summary](#) [Chromidom Report](#) [Distance tree of results](#)

[Graphic Summary](#)

[Descriptions](#)

Sequences producing significant alignments:

Select: [All](#) [Name](#) Selected: 0

[Alignments](#) [Download](#) [GenBank](#) [Graphics](#) [Distance tree of results](#)

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	Gorilla gorilla insulin precursor (INS) gene, complete cds	1375	1375	100%	0.0	100%	AF132500.1
☐	Homo sapiens insulin (INS), RefSeqGene on chromosome 11	1314	1314	99%	0.0	99%	NM_002116.1
☐	Homo sapiens chromosome 11, clone RP11-32217, complete sequence	1314	1314	99%	0.0	99%	AC132217.15
☐	Homo sapiens chromosome 11, clone RP4-539G11, complete sequence	1308	1308	99%	0.0	99%	AC132262.8
☐	Homo sapiens insulin (INS) gene, exons 1, 2, 3, and complete cds, and insulin-like growth factor (IGF2) gene, exon 1	1295	1295	99%	0.0	98%	AF128500.1
☐	Homo sapiens tyrosine hydroxylase (TH) gene, 3' and insulin (INS) gene, complete cds; insulin-like growth factor 2 (IGF2) gene, 5' and	1275	1275	99%	0.0	98%	U15446.1
☐	Pan troglodytes insulin precursor (INS) gene, complete cds	1247	1247	94%	0.0	99%	AF137497.1
☐	Human gene for proinsulin, from chromosome 11. Includes a highly polymorphic region upstream from the insulin gene, contains 1a	1245	1245	99%	0.0	97%	V00565.1
☐	Pongo abelii BAC clone CH-276-476G11 from chromosome unknown, complete sequence	1201	1201	100%	0.0	96%	AC199962.4
☐	Human insulin(IGF-II) intergenic region with adult IGF-II promoter	1197	1197	91%	0.0	99%	X05331.1
☐	Pongo pygmaeus insulin precursor (INS) gene, complete cds	1179	1179	100%	0.0	95%	AF137503.1
☐	Homo sapiens insulin-like growth factor 2, antisense (IGF2AS) gene, complete cds	1081	1081	81%	0.0	99%	AF175632.1
☐	Human insulin gene, complete cds	848	848	66%	0.0	98%	J00205.1
☐	Orangutans gene for proinsulin	243	243	10%	3e-60	98%	X51669.1
☐	Human alpha-type insulin gene and 5' flanking polymorphic region	196	196	18%	2e-45	97%	M89232.1
☐	PREDICTED: Gorilla gorilla gorilla insulin, transcript variant 2 (INS), mRNA	135	135	9%	5e-28	100%	XM_004056428.1
☐	PREDICTED: Gorilla gorilla gorilla insulin, transcript variant 1 (INS), mRNA	135	135	9%	5e-28	100%	XM_004056427.1

Figura 9.47. Coincidencias en la base de datos

A continuación, se encuentra la sección de alineamientos, verdadero corazón del informe (ver Figura 9.48).

Download ▾ GenBank Graphics ▾ Next ▴ Previous ▴ Descriptions

Gorilla gorilla insulin precursor (INS) gene, complete cds
 Sequence ID: [gbAY137500.1](#) AY137498.2 Length: 4145 Number of Matches: 1

Range 1: 1675 to 2410 [GenBank](#) [Graphics](#) ▾ Reset Match ▴ Previous Match

Score	Expect	Identity	Gaps	Strand
1375 bits(744)	0.0	744/744(100%)	0/744(0%)	Plus/Plus
Query 1	CAGGCAACCCGACACCTCTCGGCTCTCTGCAACGAGGACATGGAATAAAGCCCTGAAAC	60		
Subject 1675	CAGGCAACCCGACACCTCTCGGCTCTCTGCAACGAGGACATGGAATAAAGCCCTGAAAC	1734		
Query 61	AGCCCTGCTGTGCGCTCTGTGTCTTGGGGGCGCTGGGCGAAGCCCACTTCCCGGAC	120		
Subject 1735	AGCCCTGCTGTGCGCTCTGTGTCTTGGGGGCGCTGGGCGAAGCCCACTTCCCGGAC	1794		
Query 121	TGTTGTGAACCCCTCCGAGCTCTTCGATGCTCTTGGGTGCGACAGGTGCGACGCCA	180		
Subject 1795	TGTTGTGAACCCCTCCGAGCTCTTCGATGCTCTTGGGTGCGACAGGTGCGACGCCA	1854		
Query 181	GCGAGGCGGACATGCACTGGCTCTCCCAAGCGGGCATGCTGTGGCTGGCTGGCTAC	240		
Subject 1855	GCGAGGCGGACATGCACTGGCTCTCCCAAGCGGGCATGCTGTGGCTGGCTGGCTAC	1914		
Query 241	CCGACCCCTGTGGCTCAGGGTCCAGTATGAGAGCTCGGGGGCTCTCTGAGGGGCCAGGG	300		
Subject 1915	CCGACCCCTGTGGCTCAGGGTCCAGTATGAGAGCTCGGGGGCTCTCTGAGGGGCCAGGG	1974		
Query 301	TGGTGGGGCACTGAGAAATGACTTCTTGTTCAGTAGCTCTGGCTCTTGGAGTCCCGA	360		
Subject 1975	TGGTGGGGCACTGAGAAATGACTTCTTGTTCAGTAGCTCTGGCTCTTGGAGTCCCGA	2034		
Query 361	AGACCTGTTCAGGAAAGGGAATGAGACATTCCAGCAATTTTCCCGCCCACTAGCGCTC	420		
Subject 2035	AGACCTGTTCAGGAAAGGGAATGAGACATTCCAGCAATTTTCCCGCCCACTAGCGCTC	2094		
Query 421	CCAGGTCATATTTAGATTATTTCTGATGGAGTCCCTGCGGAGGAGGAGGCTGGGCT	480		
Subject 2095	CCAGGTCATATTTAGATTATTTCTGATGGAGTCCCTGCGGAGGAGGAGGCTGGGCT	2154		
Query 481	GAGGAGGAGGCTCTGCAAGCGGGGGCTGGGAGGTGGGAGAGCTGCTGAGAGCTA	540		
Subject 2155	GAGGAGGAGGCTCTGCAAGCGGGGGCTGGGAGGTGGGAGAGCTGCTGAGAGCTA	2214		
Query 541	CCGCTATATCCCAAGCTCTGGGCAAGCTTGGGACATGACACACCTTGGCTTCCCGGCTA	600		
Subject 2215	CCGCTATATCCCAAGCTCTGGGCAAGCTTGGGACATGACACACCTTGGCTTCCCGGCTA	2274		
Query 601	AGCTGAGCAAGGCTCTGAGGACAGACTTATGCGCAAGGCTGAGGCTGAGGCTGAGG	660		
Subject 2275	AGCTGAGCAAGGCTCTGAGGACAGACTTATGCGCAAGGCTGAGGCTGAGGCTGAGG	2334		
Query 661	GAGCCACTTGTGCTTGGAGGAGCAGGAGAGGTGAGATGGCACTATGAGTGGGCGAGG	720		
Subject 2335	GAGCCACTTGTGCTTGGAGGAGCAGGAGAGGTGAGATGGCACTATGAGTGGGCGAGG	2394		
Query 721	TGAGGGAGCGTGGGACGACCGGG 744			
Subject 2395	TGAGGGAGCGTGGGACGACCGGG 2410			

Related Information

[Gene](#) - associated gene details
[Map Viewer](#) - aligned genomic context

Figura 9.48. Alineamientos con las coincidencias en la base de datos

constante de la Bioinformática por lo que, una vez más, nos encontramos ante la disyuntiva de qué formato elegir. Aunque existen más, los formatos de alineamiento más comunes son el FASTA, el ALN, el MSF y el PIR (similar al FASTA pero con una línea adicional para incluir anotaciones).

Para la edición y publicación de alineamientos utilizaremos Jalview (<http://www.jalview.org>). Se trata de una de las herramientas más potentes que existen por la gran cantidad de funcionalidades que soporta y por su facilidad de uso. Desde la página web podemos elegir entre lanzar la aplicación completa (Jalview Desktop, que es la misma que nos podemos instalar) o una versión ejecutable desde el navegador web (Jalview Applet). En esta práctica, emplearemos la versión completa.

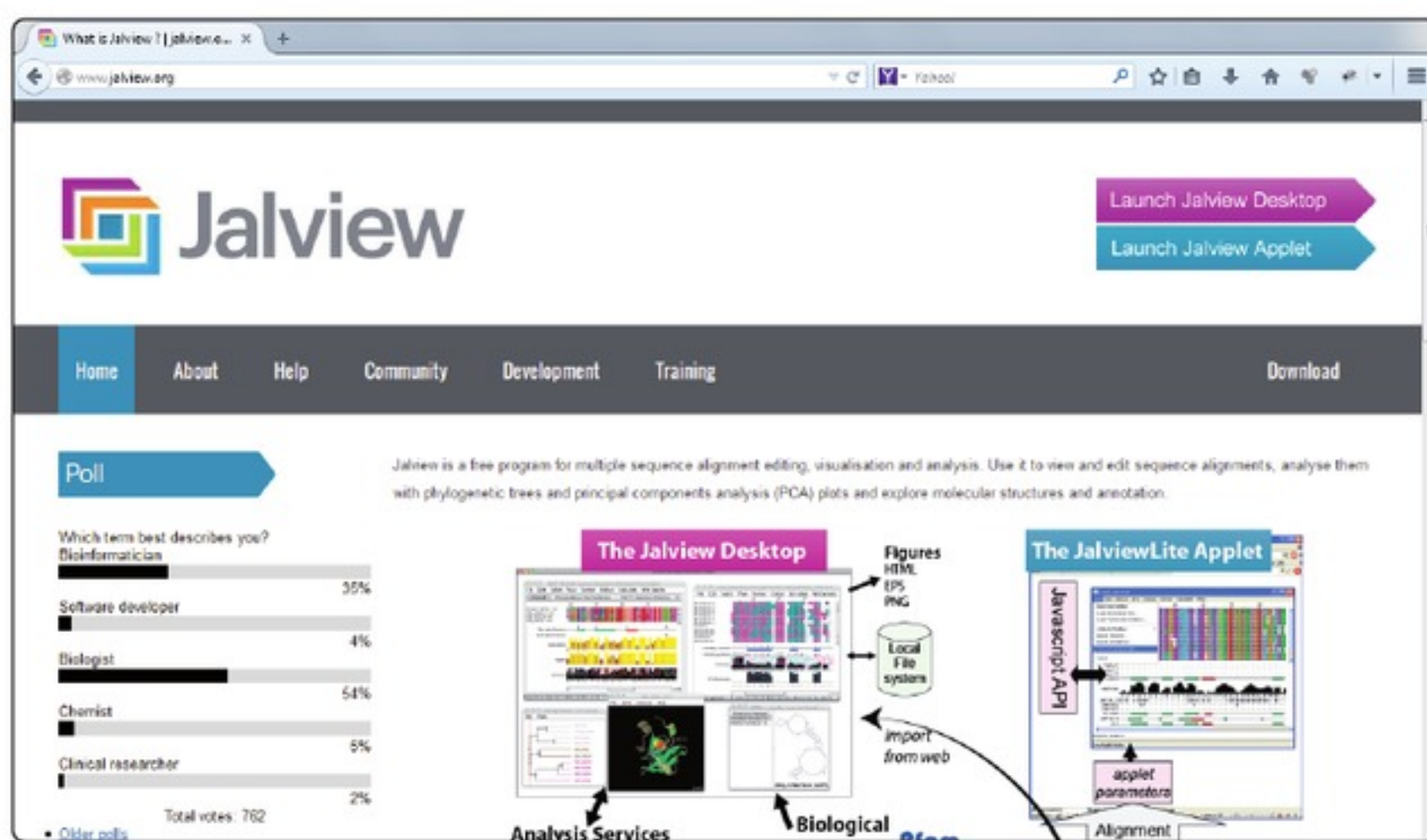


Figura 9.50. Página principal de Jalview



Aunque Jalview se utiliza *on-line* directamente desde la página web del proyecto, también es posible instalarlo en modo local. Las instrucciones están disponibles en <http://www.jalview.org>

Al pinchar en el botón **Launch Jalview Desktop**, aparece una ventana que pregunta si deseamos descargar (y ejecutar) la aplicación.

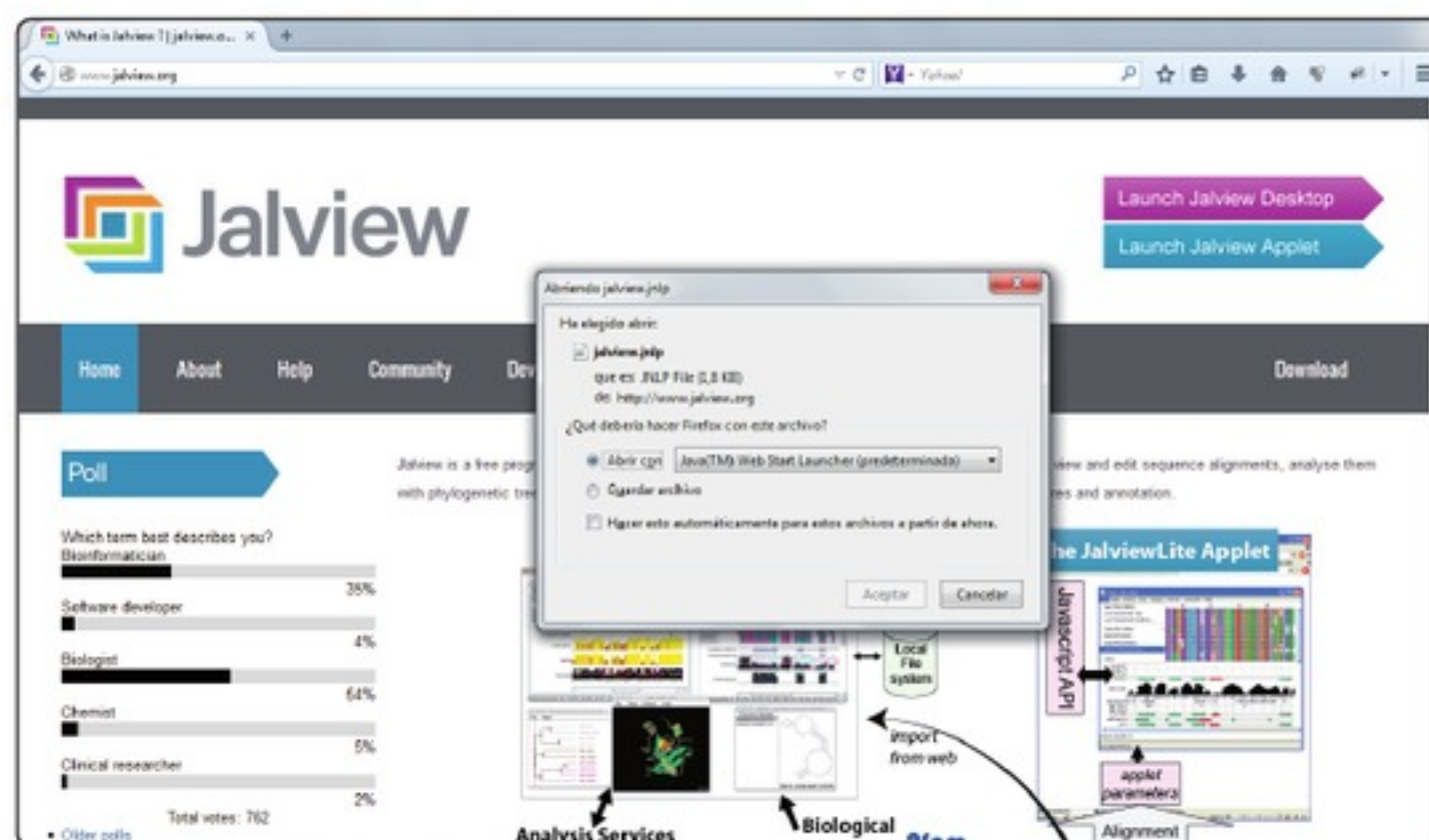


Figura 9.51. Ejecución de Jalview

Confirmamos afirmativamente y aparecerá la ventana principal de Jalview. Inicialmente, se carga un fichero con un alineamiento de ejemplo, aunque en esta práctica, por cuestiones de coherencia y continuidad, nos centraremos en el manejo de los alineamientos que hemos manejado hasta el momento.

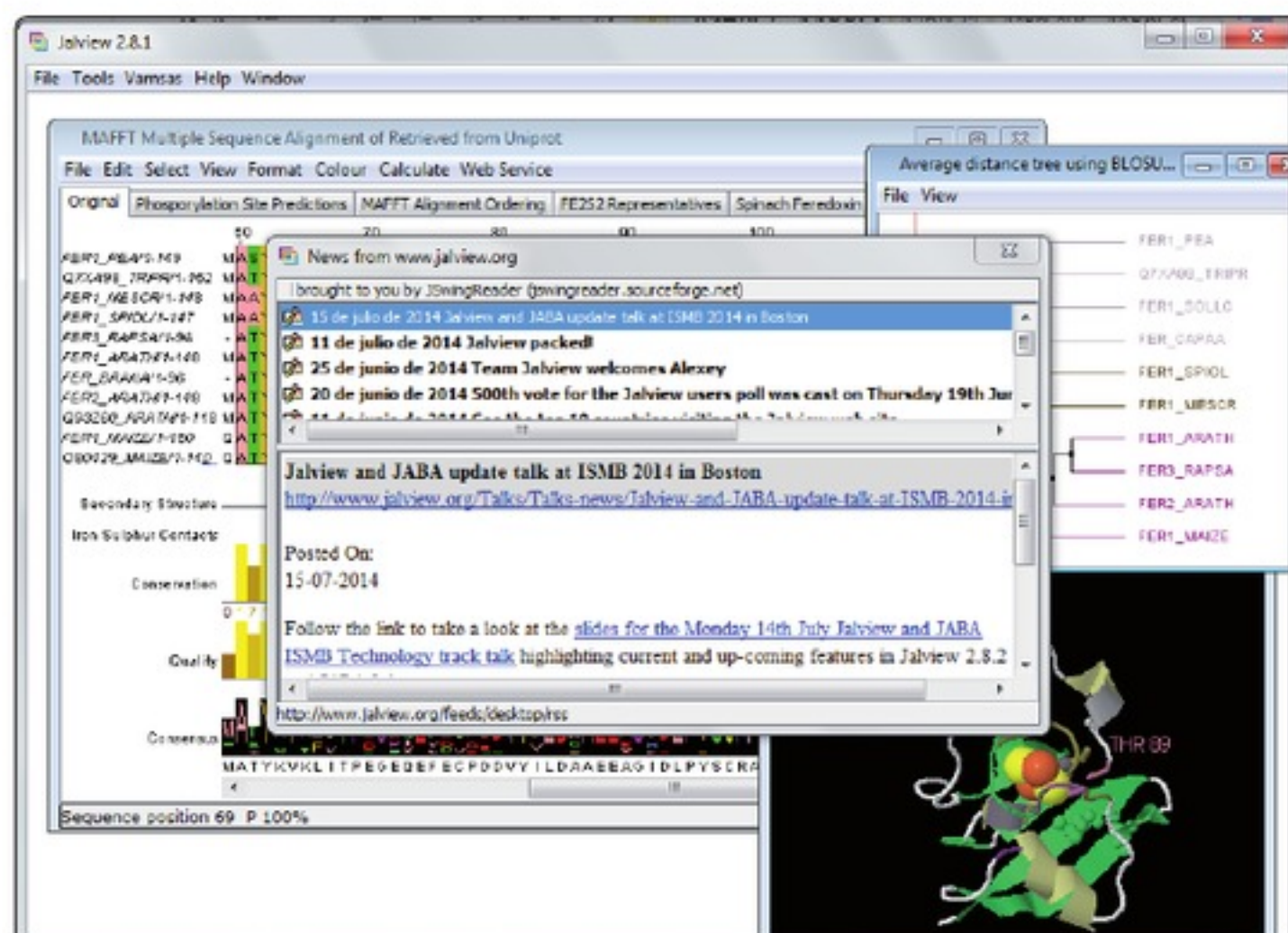


Figura 9.52. Pantalla principal de Jalview

En primer lugar, obtendremos el fichero de alineamiento generado por ClustalO para dos secuencias que, en este caso, serán las que ya utilizamos en la práctica de Dotlet y que se correspondían al ARN mensajero de Cftr de ratón (*accession* NM_021050.2) y su equivalente humano (NM_000492.3).

```
>gi|116008179|ref|NM_021050.2| Mus musculus cystic fibrosis
transmembrane conductance regulator (Cftr), mRNA
AATTGGAAGCAAATGACATCACCTCAGGTCTGAGTAAAAGGGACGAGCCAAAAGCATTGACCTGG
TCCTGGATATCCAGATGTCGAGTCCAACCTGAATTTAGCCGAACACAGACCTCATTGCCTCACGG
AGACATCATGCAGAAGTCGCCTTTGGAGAAAGCCAGCTTTATCTCCAAACTCTTCTTCAGCTGGA
CCACACCAATTTTGAGGAAAGGGTACAGACACCACTTGGAGTTGTCAGACATATAACCAAGCCCCT
TCTGCTGATTGAGCTGACCACTTGTCTGAAAACTAGAAAGAGAATGGGACAGAGAACAAGCTTC
AAAAAAGAATCCCCAGCTTATCCACGCCCTTCGGCGATGCTTTTTCTGGAGATTCCTCTTCTATG
GAATTTTGCTATACCTAGGGGAAGTCACCAAGGCTGTCCAGCCTGTCTTGCTAGGAAGAATCATA
GCATCCTATGATCCAGAAAACAAGGTGGAACGTTCCATTGCCATTTACCTTGGCATAGGCTTATG
CCTTCTCTTCATTGTCAGGACACTGCTTCTTCACCCAGCTATTTTTGGCCTTCATCGCATTGGAA
TGCAGATGAGAACAGCTATGTTTAGCTTGATTTATAAGAAGACTTTAAAGTTGTCAAGCCGCGTT
CTTGATAAAATAAGTATTGGACAACCTTGTTAGTCTTCTTTCCAACAACCTGAACAAATTTGATGA
AGGACTTGCCCTGGGCACATTTTATATGGATTGCTCCTTTACAAGTGAAGTCTTCTGATGGGGCTTC
TCTGGGACTTGTTACAGTTCTCAGCCTTCTGTGGCCTTGGTTTACTGATAATCCTGGTTATTTTT
CAAGCTATCCTAGGGAAGATGATGGTGAAGTACAGAGATCAGAGAGCTGCAAAGATCAATGAAAG
ACTCGTGATCACATCAGAAATTATTGATAATATCTATTCTGTTAAGGCATATTGTTGGGAATCAG
CGATGGAGAAAATGATTGAAAACCTGAGAGAGGTGGAGCTGAAAATGACCCGGAAGGCGGCCTAT
ATGAGGTTCTTCACTAGCTCTGCCTTCTTCTTTTCAGGGTTCTTTGTAGTCTTTCTATCTGTGCT
TCCCTACACAGTCATCAACGGAATCGTCCTACGAAAAATATTCACAACCATTTTATTCTGCATTG
TCCTACGTATGTCAGTCACACGGCAGTTCCCCACTGCCGTACAGATATGGTATGATTCTTTTGG
ATGATAAGAAAAATACAGGATTTCTTGCAGAAAACAAGAGTATAAAGTACTGGAGTATAACTTAAT
GACCACAGGCATAATCATGGAAAATGTAACAGCATTTTGGGAGGAGGGATTTGGGGAATTACTGG
AGAAAGTACAACAAGCAATGGTGACAGAAAACATTCCAGTGATGAGAACAAATGTCAGTTTCAGT
CATCTCTGCCTTGTGGGAAATCCTGTGCTGAAAAACATCAATTTGAATATAGAGAAAGGAGAGAT
GTTGGCTATTACTGGATCTACTGGATCAGGAAAGACATCACTCCTGATGTTGATTTTGGGAGAAC
TGGAAGCTTCAGAGGGAATTATTAAGCACAGTGGAAGAGTTTCATTCTGCTCTCAATTTTCTTGG
ATTATGCCGGGTACTATCAAAGAAAAATATCATCTTTGGTGTTCCTATGATGAGTACAGATATAA
GAGTGTGTCAAAGCTTGCCAACTACAGCAGGACATCACCAAGTTTGCAGAACAAGACAACACAG
TTCTTGGAGAAGGTGGAGTCACACTGAGTGGAGGTGAGCGTGCAAGGATTTCTTTAGCAAGAGCA
GTATATAAAGATGCTGATTTGTACCTATTAGATTCCCCTTTTGGATATCTAGATGTTTTTACTGA
AGAACAAGTATTTGAAAGCTGTGTTTGTAAATTGATGGCCAACAAAACCTAGGATTTTGGTTACAT
CTAAAATGGAACACTTAAGGAAAGCTGACAAAATACTAATTTTGCATCAGGGCAGTAGCTATTTT
TATGGGACATTTTCTGAGCTACAAAGTCTACGTCCAGACTTCAGTTCGAAACTCATGGGGTATGA
TACTTTTGACCAGTTTACTGAGGAAAGAAGAAGTTCAATTCTAACTGAGACCTTACGCAGGTTCT
CAGTAGACGATTCCCTCTGCCCCGTGGAGCAAACCCAAACAGTCGTTTACAGAGACTGGAGAGGTG
GGAGAAAAAAGGAAGAACTCTATTCTAAATTCATTGAGCTCTGTAAGGAAAATTTCCATTGTGCA
AAAGACTCCATTATGTATCGATGGAGAGTCTGATGATCTCCAAGAAAAGAGACTGTCCCTAGTTC
CGGATTCTGAACAGGGGGAGGCTGCTCTGCCGCGCAGCAACATGATCGCCACCGGCCCCACATTT
CCAGGCAGAAGAAGACAGTCTGTTTTGGATCTGATGACGTTACACCCAACTCAGGCTCCAGCAA
TCTTCAGAGGACCAGAACTTCTATTGAAAAATCTCCTTAGTCCCTCAGATAAGCTTAAATGAAG
```


TGGATGTATATTCAAGGAGATTATCGCAAGATAGCACACTGAACATCACTGAAGAAATTAACGAA
GAAGATTTAAAGGAGTGTTTTCTTGATGATGTGATCAAGATACCCCCGGTGACAACATGGAACAC
ATACCTACGATATTTTACTCTCCATAAAGGCTTACTGCTAGTGCTGATTTGGTGCGTACTGGTTT
TTCTGGTTGAGGTGGCTGCTTCTTTATTTGTGTTATGGTTGCTTAAAAACAACCCTGTTAACAGT
GGAAACAATGGTACTAAAATTTCCAATAGCTCCTATGTTGTGATCATCACCAGTACCAGTTTCTA
TTATATTTTTTACATTTACGTGGGAGTGGCTGACACTTTGCTTGCCCTGAGCCTCTTCAGAGGTT
TGCCGCTGGTGACATCGTTAATCACAGCATCAAAAATTTTGACAGGAAAATGTTACACTCCATT
CTTCACGCCCCCTATGTCGACCATCAGCAAGCTGAAAGCAGGTGGGATTCTTAACAGATTCTCCAA
AGATATAGCAATTTTGGATGACTTTCTGCCTCTTACCATTTTTGACTTCATTTCAGTTGGTGTTCA
TTGTGATTGGAGCTATAATAGTCGTCTCGGCATTACAACCCTACATCTTCCTAGCAACGGTGCCA
GGGCTAGTAGTCTTTATTTTACTGAGGGCCTACTTCCTTCATACAGCACAGCAGCTCAAACAAC
GGAATCTGAAGGCAGGAGTCCAATTTTCACCCACCTTGTGACAAGCTTAAAAGGACTCTGGACAC
TTCGAGCCTTCCGACGCCAGACTTACTTTGAAACTCTGTTCCACAAAGCTCTGAATTTGCACACT
GCCAACTGGTTTATGTATCTGGCAACCTTGCGCTGGTTCCAAATGAGAATAGACATGATATTTGT
CCTCTTCTTCATTGTTGTTACCTTCATCTCCATTTTAACAACAGGTGAAGGAGAAGGAACAGCTG
GTATTATTCTAACTTTAGCTATGAATATCATGAGTACTTTGCAGTGGGCTGTGAACCTCAAGCATT
GATACAGATAGCTTGATGCGATCTGTGAGCAGAGTGTTTAAGTTTATTGATATACAAACAGAAGA
AAGTATGTACACACAGATAATTAAAGAACTACCTAGAGAAGGATCATCTGACGTTTTAGTCATTA
AGAATGAGCATGTGAAGAAAAGTGATATCTGGCCCTCTGGAGGCGAAATGGTTGTCAAAGACCTT
ACTGTGAAATACATGGATGATGGAAATGCCGTATTAGAGAACATTTCTTTTTCAATAAGTCCTGG
ACAGAGGGTGGGGCTCTTAGGAAGAACTGGATCAGGAAAAAGTACTTTGCTTTCAGCATTTTTAC
GAATGTTGAACATTAAAGGTGATATAGAGATTGATGGTGTCTCATGGAATTCAGTGACCTTACAA
GAATGGAGGAAAGCTTTCGGAGTGATAACACAGAAAGTATTTATCTTTTCTGGAACATTCAGACA
AAACCTGGATCCCAATGGAAAATGGAAAGATGAAGAAATATGGAAAGTTGCAGATGAGGTTGGAC
TCAAGTCTGTAATAGAGCAGTTTCCTGGACAGCTCAACTTTACCCTTGTGGATGGGGGTTATGTG
CTAAGCCATGGCCATAAGCAATTAATGTGCTTGGCCCGATCAGTTCTCAGTAAGGCCAAGATCAT
ACTGCTTGATGAGCCCAGTGCCCATCTAGACCCCATACATAACCAAGTCATTCGACGAGTTCTAA
AACAAGCCTTCGCTGGTTGCACAGTCATCCTCTGTGAACACAGGATAGAAGCGATGTTGGATTGC
CAGCGATTTTTGGTCATAGAAGAGAGCAATGTCTGGCAGTACGACTCCCTTCAGGCACTTCTGAG
TGAGAAGAGTATCTTCCAGCAGGCCATTAGCTCCTCGGAAAAGATGAGGTTCTTCCAGGGCCGCC
ACTCCAGCAAGCACAAAGCCTCGGACGCAAATTAAGTCTGCTGAAAGAGGAGACAGAAGAAGAAGTT
CAAGAAACCCGTCTCTAGTGCTGGGATGCTGAGGAAGCAACTCAGTGCCTGAGTCCATTCCCAG
AACCCATGCAGAATGAAAAAAGCCAGGCATTTCCCATGCTTCTAACCCAGTGCTGGGGACACAG
AGACAGGTGGATCCCTGGGGCTCTGTGGCAAGTGATCCTAGCCACAAAGAGAGTTCCAGGCTGG
GCACCTGAGGGACAATACCTGTGGATATACTCTTGCTTCCACATGCAAGTACATATACACATGCA
TGCACATTAGTGGACATACACAGAAAAGCAAAGAAGAAGGAAAGAGGGAAGAAAATAGTGCAA
ATAATTGCAAAACGATCATGTATGGAGTCTGCTCATGGACTTAGAGGAGGTGAACTCTACTACCT
GTGCCTTTGAAAGAAGGGTGAAGCCTGCGACTTGCTCTTTAAGAGACTGTTTTGGAAGAGAGTTC
AAAAACGTTTCATATGGGTATGGGTAACTGACTTTCCAGCAGTAGTCAAATTGTTTGAACCTTCAGA
TAGTTGATAATGACCACTTGTGTATTGCAAGGCAGATTTTTCTGAAAACATTTGCCCCCTAATAG
TAGCTGAAAAAGCAGCTATAAATGCCAACCAGGTTAGTCATTCGGCTTATTGTTTCAGTACAGCTG
GTTAATTTGCATTATTGAAGAACTGAAATTATAGTGCTTAGATATAGGACAAAGTAAAGAGAAGT
AAAAACAGTGTCTTATATACTCAAAGCCCACTTACTTTCTCTAAGATATGTATTGCCTTCTA
TACATTGTCTGCCCCATTCCAAGCAAATGTTAGAATATTATACAAAATACTGGGTGGTATTGATT
GAAAGATGCCCCGACATCTGGTGATCTAGTAACCCATCAGGATTAAGGATATCCAGGTCTTGGA
TTAAGGTAAAGACCATCTAGCCTTACTACCGTACAGCTAAACATTCTTATTACCAGAATAAGACC
TAGGAAAAGAACTGTTTCAGTCCCATAAAGTGGCCTGGATAATTTCTTGGATATGGAAATCGACA

CACTTATGTTCCCAGAAAGCAACAGATCTTTAAGACTTCTGAAGTGAAGGAAGGTTGTGTTAGTG
CAAAC TAGTGCAGCCCAGTGCCAGGTCCAGGAGTTAACATGTAGACAGGCCATGGACTGTGTGGG
TAGATGCTCATGGAAATGTGCAGTAGTATGTTTCATGTGCTCTCAGCTAGCTGTGTGTACTTCAA
CTGTCTCCACAGAGTTGTTGGGGAGACACTCTGAAAAAGAATTAATTGTGAATTAGTTTTATATA
CTTTGTTTTATAATTTGTGATGCAAATGAAAATTTCTCTGGGAAATATTTATTTTAGTAATAATG
TTTCAAAC TCATATATAACAATGCTGTATTTTAAGAATGATTACATAATGACTTATATTTGTATA
AAATAATTTTTATATTTGAAATGTTAACTTTTTATAGCACTAGCTATTTTAAAACAGGGGAGTGA
GGAGGACAGGGATGATAAGGATCATTCAACTTCATGTTGTGAAGACGAGCTGATGTAAATCTTGT
ACCCATCTGTGTGGTTCTCAGACAACACATGCTCTCTTTTAATGCAGCTTTGAAGAAGATGGTAC
CAAAGGTTAAGACGGCCCCCTGATGGGCACATCAACTTCTGAACTGCAAAC TAAGCTTTAGAGGA
ATGTATTATATTTATTACTGTAATAGAATATCATGTGTCAATAAAATCCTTTTATTTGTGTGAAA
>gi|90421312|ref|NM_000492.3| Homo sapiens cystic fibrosis
transmembrane conductance regulator (ATP-binding cassette
sub-family C, member 7) (CFTR), mRNA
AATTGGAAGCAAATGACATCACAGCAGGTCAGAGAAAAAGGGTTGAGCGGCAGGCACCCAGAGTA
GTAGGTCTTTGGCATTAGGAGCTTGAGCCCAGACGGCCCTAGCAGGGACCCAGCGCCCGAGAGA
CCATGCAGAGGTCGCCTCTGGAAGAGGCCAGCGTTGTCTCCAACTTTTTTTTCAGCTGGACCAGA
CCAATTTTGAGGAAAGGATACAGACAGCGCCTGGAATTGTCAGACATATACCAAATCCCTTCTGT
TGATTCTGCTGACAATCTATCTGAAAAATTGGAAAGAGAATGGGATAGAGAGCTGGCTTCAAAGA
AAAATCCTAACTCATTAATGCCCTTCGGCGATGTTTTTTCTGGAGATTTATGTTCTATGGAATC
TTTTTATATTTAGGGGAAGTCACCAAAGCAGTACAGCCTCTCTTACTGGGAAGAATCATAGCTTC
CTATGACCCGGATAACAAGGAGGAACGCTCTATCGCGATTTATCTAGGCATAGGCTTATGCCTTC
TCTTTATTGTGAGGACACTGCTCCTACACCCAGCCATTTTTGGCCTTCATCACATTGGAATGCAG
ATGAGAATAGCTATGTTTAGTTTGATTTATAAGAAGACTTTAAAGCTGTCAAGCCGTGTTCTAGA
TAAAATAAGTATTGGACAACCTGTTAGTCTCCTTTCCAACAACCTGAACAAATTTGATGAAGGAC
TTGCATTGGCACATTTTCGTGTGGATCGCTCCTTTGCAAGTGGCACTCCTCATGGGGCTAATCTGG
GAGTTGTTACAGGCGTCTGCCTTCTGTGGACTTGGTTTCCTGATAGTCCTTGCCCTTTTTTCAGGC
TGGGCTAGGGAGAATGATGATGAAGTACAGAGATCAGAGAGCTGGGAAGATCAGTGAAAGACTTG
TGATTACCTCAGAAATGATTGAAAATATCCAATCTGTTAAGGCATACTGCTGGGAAGAAGCAATG
GAAAAAATGATTGAAAACCTTAAGACAAACAGAACTGAACTGACTCGGAAGGCAGCCTATGTGAG
ATACTTCAATAGCTCAGCCTTCTTCTCTCAGGGTTCTTTGTGGTGTTTTTATCTGTGCTTCCCT
ATGCACTAATCAAAGGAATCATCCTCCGGAAAATATTCACCACCATCTCATTCTGCATTGTTCTG
CGCATGGCGGTCACTCGGCAATTTCCCTGGGCTGTACAAACATGGTATGACTCTCTTGGAGCAAT
AAACAAAATACAGGATTTCTTACAAAAGCAAGAATATAAGACATTGGAATATACTTAACGACTA
CAGAAGTAGTGATGGAGAATGTAACAGCCTTCTGGGAGGAGGGATTTGGGGAATTATTTGAGAAA
GCAAAACAAAACAATAACAATAGAAAACTTCTAATGGTGATGACAGCCTCTTCTTCAGTAATTT
CTCACTTCTTGGTACTCCTGTCCTGAAAGATATTAATTTCAAGATAGAAAGAGGACAGTTGTTGG
CGGTTGCTGGATCCACTGGAGCAGGCAAGACTTCACTTCTAATGGTGATTATGGGAGAACTGGAG
CCTTCAGAGGGTAAAATTAAGCACAGTGGAAGAATTTCAATTCTGTTCTCAGTTTTCTGGATTAT
GCCTGGCACCATTAAAGAAAATATCATCTTTGGTGTTTCCTATGATGAATATAGATACAGAAGCG
TCATCAAAGCATGCCAACTAGAAGAGGACATCTCCAAGTTTGCAGAGAAAGACAATATAGTTCTT
GGAGAAGGTGGAATCACACTGAGTGGAGGTCAACGAGCAAGAATTTCTTTAGCAAGAGCAGTATA
CAAAGATGCTGATTTGTATTTATTAGACTCTCCTTTTGGATACCTAGATGTTTTAACAGAAAAAG
AAATATTTGAAAGCTGTGTCTGTAACTGATGGCTAACAAAAC TAGGATTTTGGTCACTTCTAAA
ATGGAACATTTAAAGAAAGCTGACAAAATATTAATTTTGCATGAAGGTAGCAGCTATTTTTATGG
GACATTTTCAGAACTCCAAAATCTACAGCCAGACTTTAGCTCAAACTCATGGGATGTGATTCTT
TCGACCAATTTAGTGCAGAAAGAAGAAATTCATCCTAACTGAGACCTTACACCGTTTCTCATTA

GAAGGAGATGCTCCTGTCTCCTGGACAGAAACAAAAACAATCTTTTAAACAGACTGGAGAGTT
TGGGGAAAAAAGGAAGAATTCTATTCTCAATCCAATCAACTCTATACGAAAATTTTCCATTGTGC
AAAAGACTCCCTTACAAATGAATGGCATCGAAGAGGATTCTGATGAGCCTTTAGAGAGAAGGCTG
TCCTTAGTACCAGATTCTGAGCAGGGAGAGGCGATACTGCCTCGCATCAGCGTGATCAGCACTGG
CCCCACGCTTCAGGCACGAAGGAGGCAGTCTGTCCTGAACCTGATGACACACTCAGTTAACCAAG
GTCAGAACATTACCGAAAGACAACAGCATCCACACGAAAAGTGTCCTGGCCCCTCAGGCAAAC
TTGACTGAACTGGATATATATTCAAGAAGGTTATCTCAAGAACTGGCTTGGAATAAGTGAAGA
AATTAACGAAGAAGACTTAAAGGAGTGCTTTTTTGTGATGATATGGAGAGCATACCAGCAGTGACTA
CATGGAACACATACCTTCGATATATTACTGTCCACAAGAGCTTAATTTTTGTGCTAATTTGGTGC
TTAGTAATTTTTCTGGCAGAGGTGGCTGCTTCTTTGGTTGTGCTGTGGCTCCTTGGAACACTCC
TCTTCAAGACAAAGGGAATAGTACTCATAGTAGAAATAACAGCTATGCAGTGATTATCACCAGCA
CCAGTTCGTATTATGTGTTTTACATTTACGTGGGAGTAGCCGACACTTTGCTTGCTATGGGATTC
TTCAGAGGTCTACCACTGGTGCATACTCTAATCACAGTGTGCGAAAATTTTACACCACAAAATGTT
ACATTCTGTTCTTCAAGCACCTATGTCAACCCTCAACACGTTGAAAGCAGGTGGGATTCTTAATA
GATTCTCCAAAGATATAGCAATTTTGGATGACCTTCTGCCTCTTACCATATTTGACTTCATCCAG
TTGTTATTAATTGTGATTGGAGCTATAGCAGTTGTGCGAGTTTTACAACCCTACATCTTTGTTGC
AACAGTGCCAGTGATAGTGGCTTTTATTATGTTGAGAGCATATTTCTCCAAACCTCACAGCAAC
TCAAACAACCTGGAATCTGAAGGCAGGAGTCCAATTTTCACTCATCTTGTTACAAGCTTAAAAGGA
CTATGGACACTTCGTGCCTTCGGACGGCAGCCTTACTTTGAACTCTGTTCCACAAAGCTCTGAA
TTTACATACTGCCAACTGGTTCTTGTTACCTGTCAACACTGCGCTGGTTCCAAATGAGAATAGAAA
TGATTTTTGTGATCTTCTTCATTGCTGTTACCTTCATTTCCATTTTAACAACAGGAGAAGGAGAA
GGAAGAGTTGGTATTATCCTGACTTTAGCCATGAATATCATGAGTACATTGCAGTGGGCTGTAAA
CTCCAGCATAGATGTGGATAGCTTGATGCGATCTGTGAGCCGAGTCTTTAAGTTCATTGACATGC
CAACAGAAGGTAAACCTACCAAGTCAACCAAACCATACAAGAATGGCCAACTCTCGAAAGTTATG
ATTATTGAGAATTCACACGTGAAGAAAGATGACATCTGGCCCTCAGGGGGCCAAATGACTGTCAA
AGATCTCACAGCAAAATACACAGAAGGTGGAAATGCCATATTAGAGAACATTTCTTCTCAATAA
GTCCTGGCCAGAGGGTGGGCCTCTTGGGAAGAACTGGATCAGGGAAGAGTACTTTGTTATCAGCT
TTTTTGAGACTACTGAACACTGAAGGAGAAATCCAGATCGATGGTGTGTCTTGGGATTCAATAAC
TTTGCAACAGTGGAGGAAAGCCTTTGGAGTGATACCACAGAAAGTATTTATTTTTCTGGAACAT
TTAGAAAAAATTTGGATCCCTATGAACAGTGGAGTGATCAAGAAATATGGAAAGTTGCAGATGAG
GTTGGGCTCAGATCTGTGATAGAACAGTTTCTGGGAAGCTTGACTTTGTCCTTGTGGATGGGGG
CTGTGTCCTAAGCCATGGCCACAAGCAGTTGATGTGCTTGGCTAGATCTGTTCTCAGTAAGGCGA
AGATCTTGCTGCTTGATGAACCCAGTGCTCATTTGGATCCAGTAACATACCAAATAATTAGAAGA
ACTCTAAAACAAGCATTTGCTGATTGCACAGTAATCTCTGTGAACACAGGATAGAAGCAATGCT
GGAATGCCAACAATTTTTGGTCATAGAAGAGAACAAGTGCGGCAGTACGATTCCATCCAGAAAC
TGCTGAACGAGAGGAGCCTCTTCCGGCAAGCCATCAGCCCCCTCCGACAGGGTGAAGCTCTTTCCC
CACCGGAACTCAAGCAAGTGCAAGTCTAAGCCCCAGATTGCTGCTCTGAAAGAGGAGACAGAAGA
AGAGGTGCAAGATACAAGGCTTTAGAGAGCAGCATAAATGTTGACATGGGACATTTGCTCATGGA
ATTGGAGCTCGTGGGACAGTCACCTCATGGAATTGGAGCTCGTGGAACAGTTACCTCTGCCTCAG
AAAACAAGGATGAATTAAGTTTTTTTTTAAAAAAGAAACATTTGGTAAGGGGAATTGAGGACACT
GATATGGGTCTTGATAAATGGCTTCTGGCAATAGTCAAATTTGTGTGAAAGGTAATTCAAATCCT
TGAAGATTTACCACTTGTTGTTTTGCAAGCCAGATTTTCTGAAAACCCTTGCCATGTGCTAGTAA
TTGGAAAGGCAGCTCTAAATGTCAATCAGCCTAGTTGATCAGCTTATTGTCTAGTGAACTCGTT
AATTTGTAGTGTTGGAGAAGAACTGAAATCATACTTCTTAGGGTTATGATTAAGTAATGATAACT
GGAAACTTCAGCGGTTTATATAAGCTTGTATTCCTTTTTCTCTCCTCTCCCCATGATGTTTAGAA
ACACAACATATATTGTTTGCTAAGCATTCCAACATCTCATTTCCAAGCAAGTATTAGAATACCAC
AGGAACCACAAGACTGCACATCAAAATATGCCCCATTCAACATCTAGTGAGCAGTCAGGAAAGAG


```

AACTTCCAGATCCTGGAAATCAGGGTTAGTATTGTCCAGGTCTACCAAAAATCTCAATATTTTCAG
ATAATCACAATACATCCCTTACCTGGGAAAGGGCTGTTATAATCTTTCACAGGGGACAGGATGGT
TCCCTTGATGAAGAAGTTGATATGCCTTTTCCCAACTCCAGAAAGTGACAAGCTCACAGACCTTT
GAACTAGAGTTTAGCTGGAAAAGTATGTTAGTGCAAATTGTACAGGACAGCCCTTCTTTCCACA
GAAGCTCCAGGTAGAGGGTGTGTAAGTAGATAGGCCATGGGCACTGTGGGTAGACACACATGAAG
TCCAAGCATTTAGATGTATAGGTTGATGGTGGTATGTTTTTCAGGCTAGATGTATGTACTTCATGC
TGTCTACACTAAGAGAGAATGAGAGACACACTGAAGAAGCACCATCATGAATTAGTTTTATATG
CTTCTGTTTTATAATTTTGTGAAGCAAAATTTTTCTCTAGGAAATATTTATTTTAATAATGTTT
CAAACATATATAACAATGCTGTATTTTAAAAGAATGATTATGAATTACATTTGTATAAAATAATT
TTTATATTTGAAATATTGACTTTTTATGGCACTAGTATTTCTATGAAATATTATGTTAAACTGG
GACAGGGGAGAACCTAGGGTGATATTAACCAGGGGGCCATGAATCACCTTTTGGTCTGGAGGGAAG
CCTTGGGGCTGATGCAGTTGTTGCCACAGCTGTATGATTCCCAGCCAGCACAGCCTCTTAGATG
CAGTTCTGAAGAAGATGGTACCACCAGTCTGACTGTTTCCATCAAGGGTACACTGCCTTCTCAAC
TCCAAACTGACTCTTAAGAAGACTGCATTATATTTATTACTGTAAGAAAATATCACTTGTCAATA
AAATCCATACATTTGTGTGAAA

```

Figura 9.53. Secuencias bajo estudio

Una vez generado el fichero de alineamiento, lo descargaremos y lo guardaremos en local.

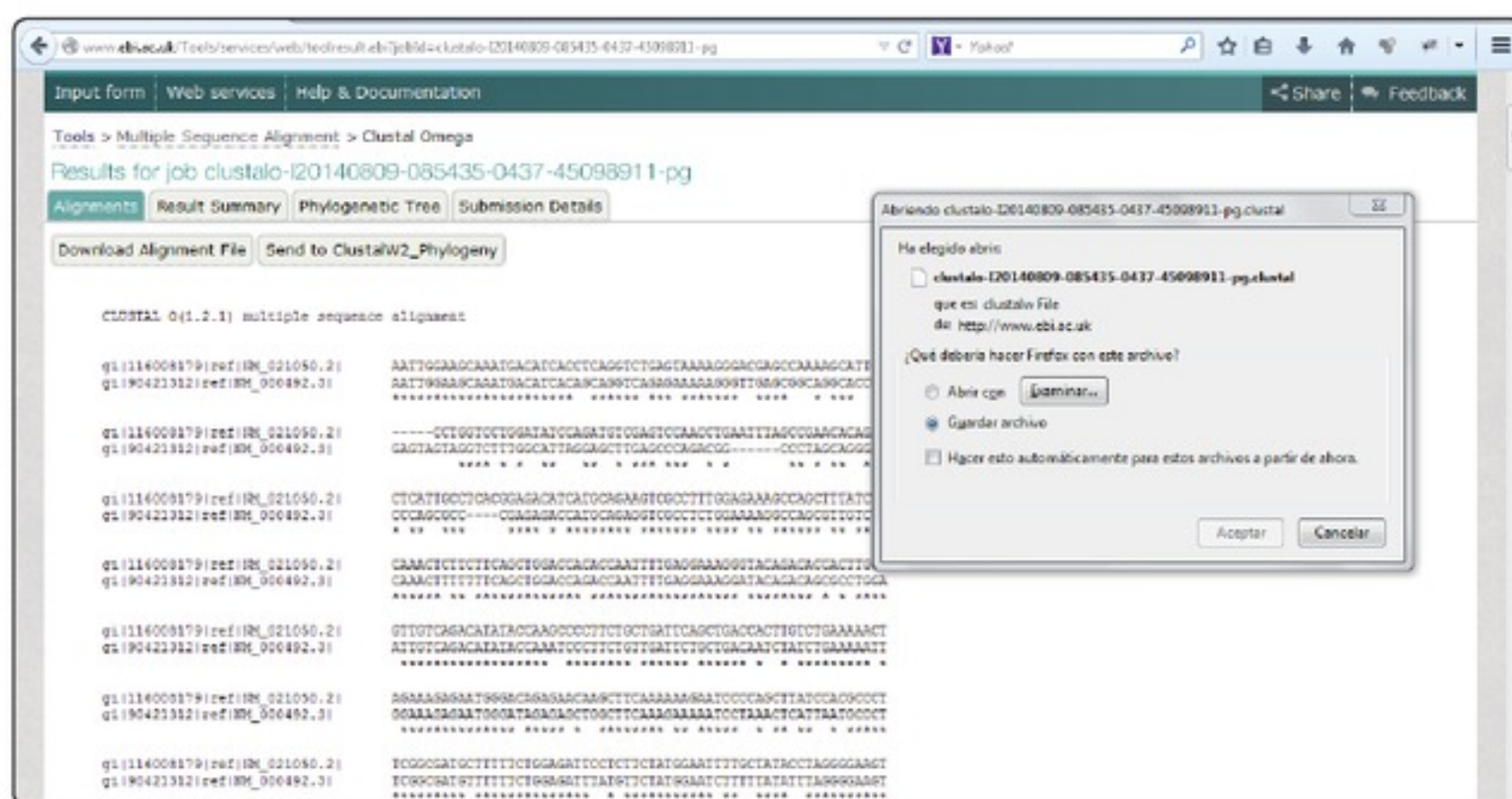


Figura 9.54. Descarga del fichero de alineamiento

A continuación, en el menú **Archivo** de Jalview, seleccionamos la opción de cargar el fichero desde un fichero local (el que acabamos de guardar):

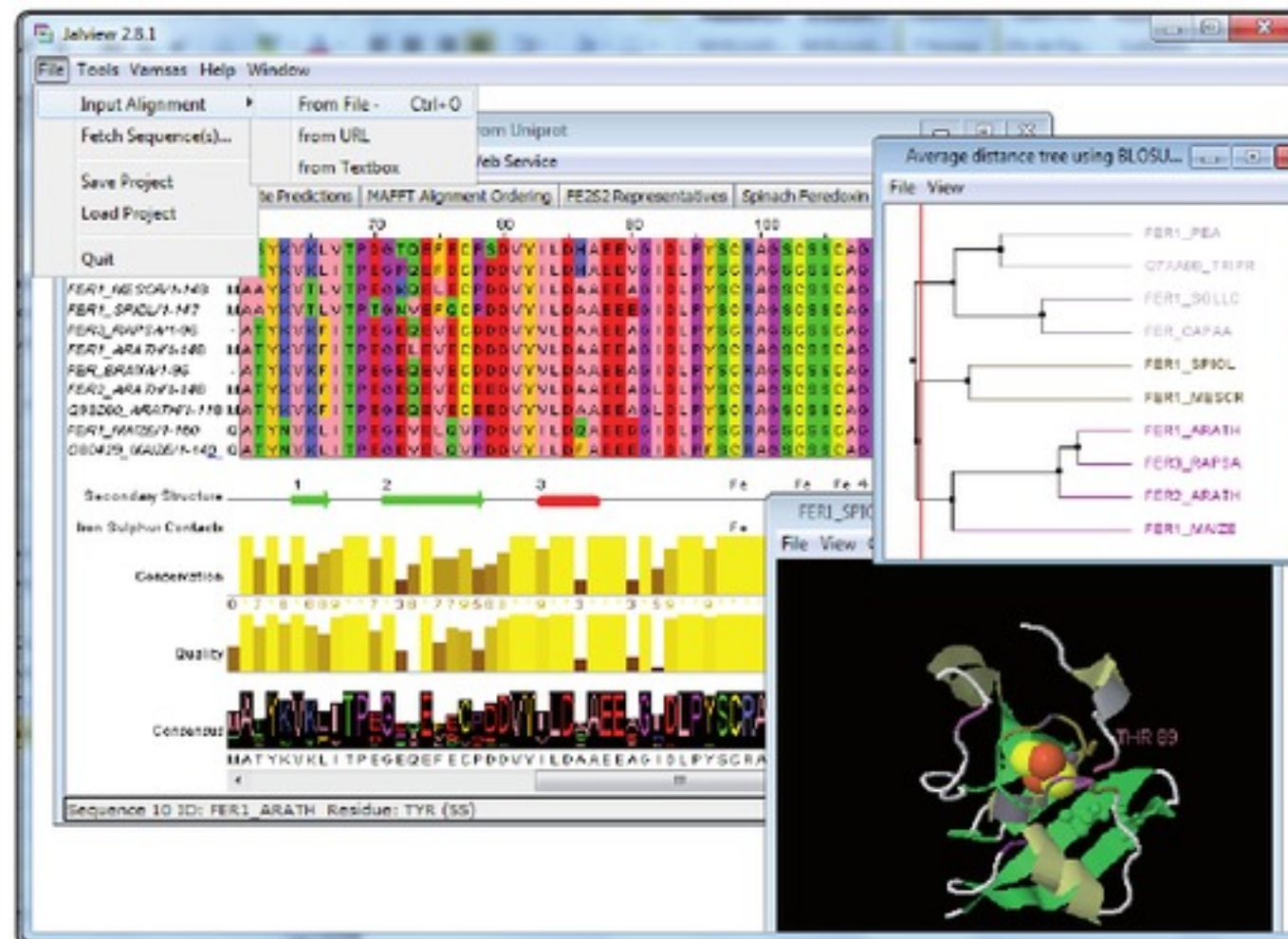


Figura 9.55. Carga de un fichero local en Jalview

Una vez seleccionado el fichero, Jalview mostrará una representación gráfica del alineamiento múltiple (ver Figura 9.56):



Figura 9.56. Representación gráfica del alineamiento

En la opción **Colour**, de todos los modelos cromáticos soportados, seleccionamos la opción **Clustalx**, ya que colorear el alineamiento simplifica la inspección visual del mismo (ver Figura 9.57).

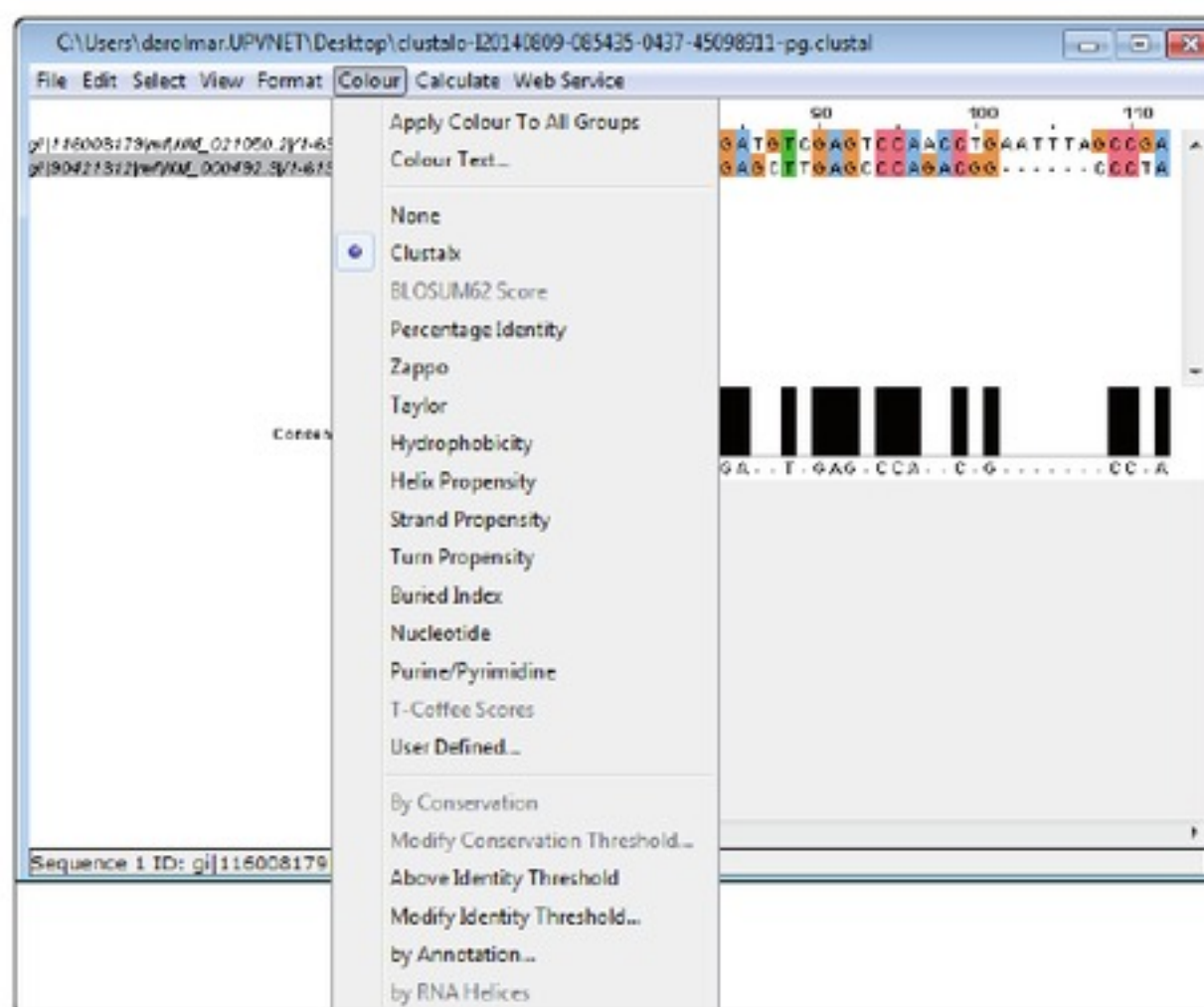


Figura 9.57. Cambio del modelo cromático



Cambia los valores del modelo cromático y comprueba cómo varía la representación gráfica de la secuencia.

9.4.1 Creación de grupos

Un grupo es una región etiquetada de un alineamiento que puede manejarse de manera autónoma para, por ejemplo, seleccionar dicha región directamente sin tener que volver a buscar los fragmentos de las secuencias implicadas.

Para ello, se seleccionan las secuencias mientras se mantiene pulsada la tecla **Ctrl**. En este caso, vamos a crear un grupo con los nucleótidos comprendidos entre la posición 68 y la 129 (ver Figura 9.58). Una vez seleccionadas las secuencias, pulsamos el botón derecho del ratón y en el menú desplegable, indicamos **Select ▶ Create**

Groups. Cuando se ha creado un grupo, las secuencias aparecen recuadradas en rojo con trazo discontinuo en la parte superior (ver Figura 9.58). A partir de ese momento, el grupo se comportará, a efecto de la edición del alineamiento, como una unidad.

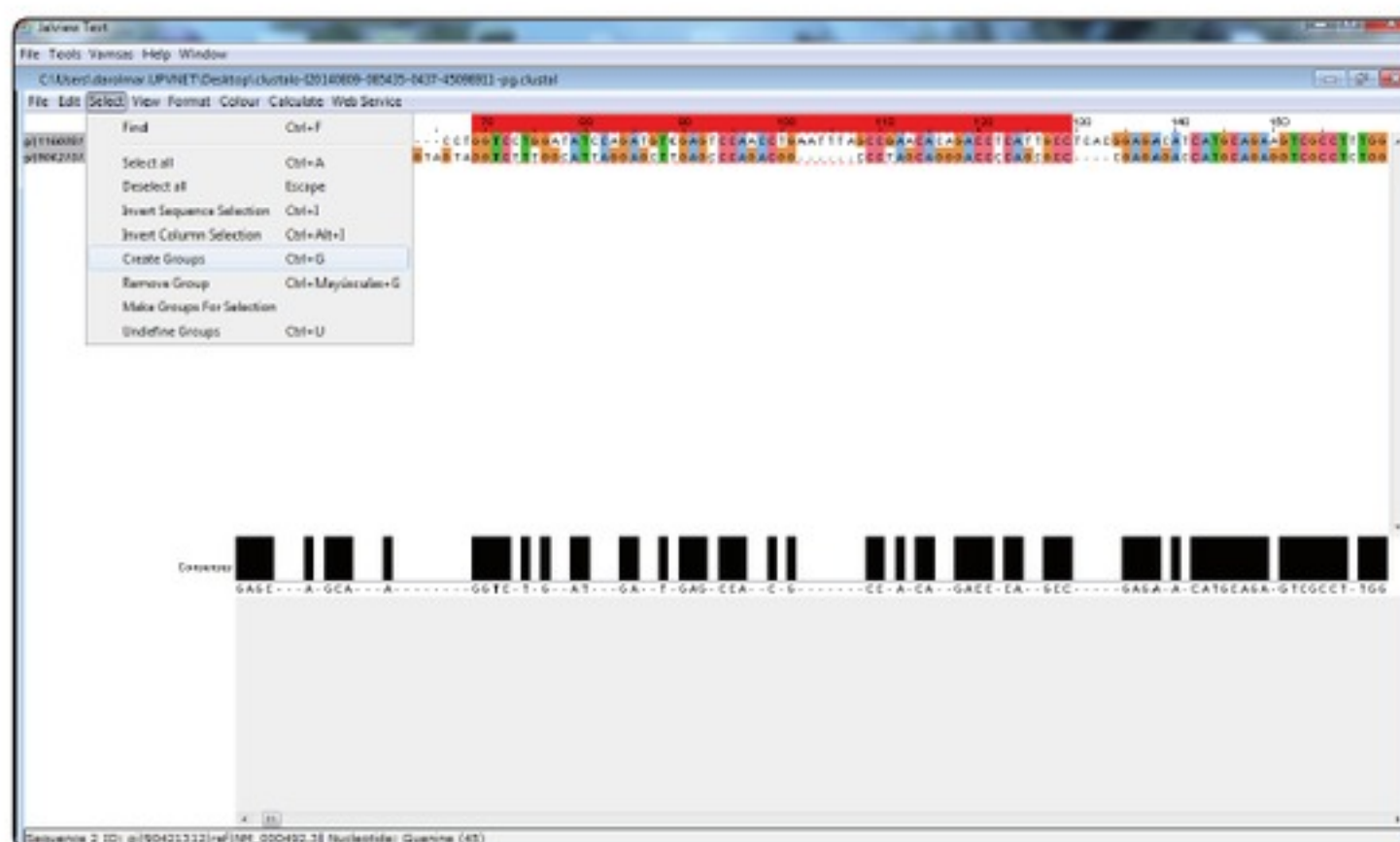


Figura 9.58. Creación de un grupo con Jalview

Es posible editar el grupo para darle un nombre descriptivo, pinchando, sobre la selección del grupo, el botón derecho del ratón. Aparecerá un menú desplegable con las opciones descritas en la Figura 9.59:

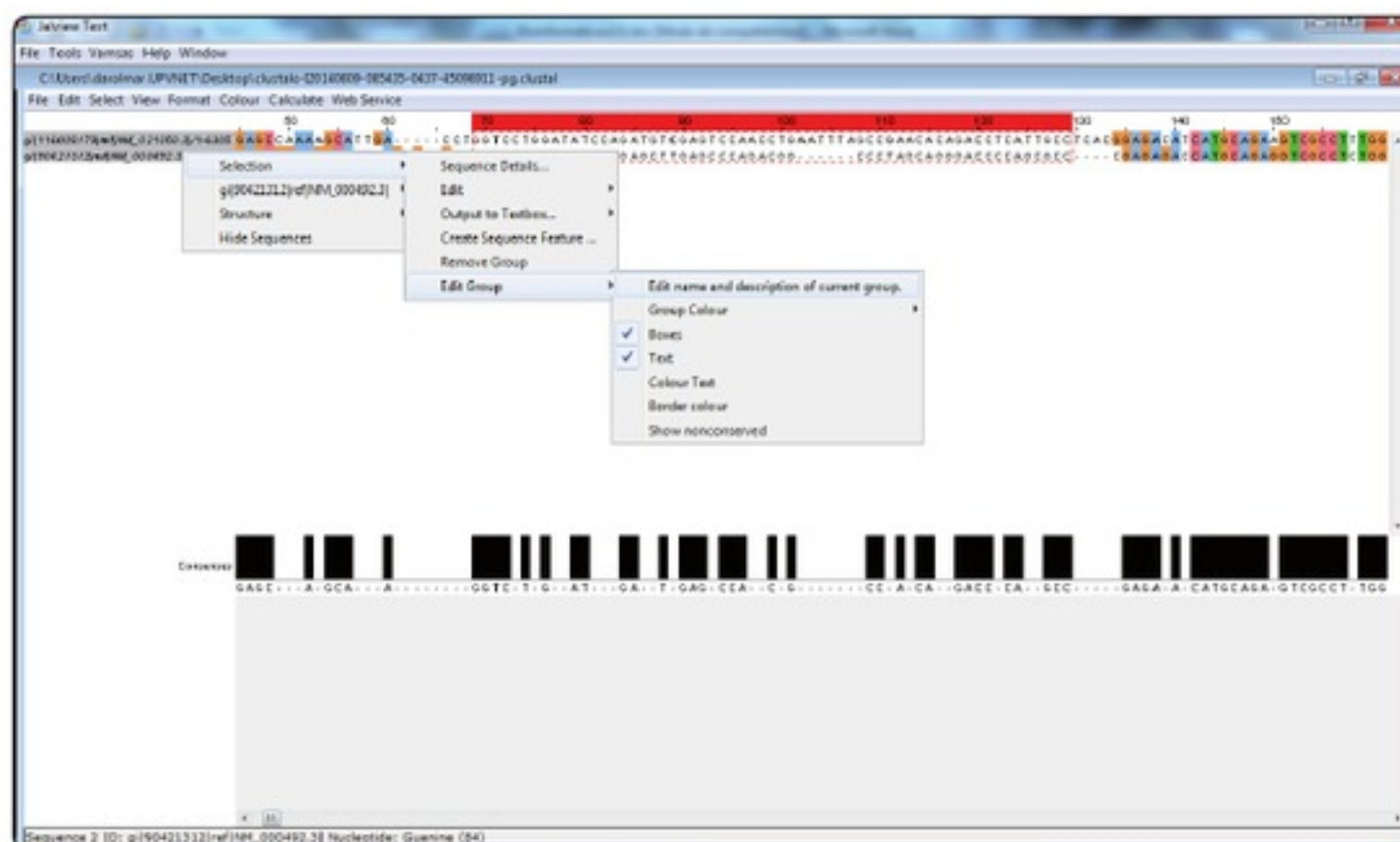


Figura 9.59. Edición de un grupo

Si seleccionamos la opción **Edit name and description of current group**, aparecerá una caja de diálogo en donde podremos introducir el nombre y la descripción del grupo que hemos seleccionado (ver Figura 9.60).

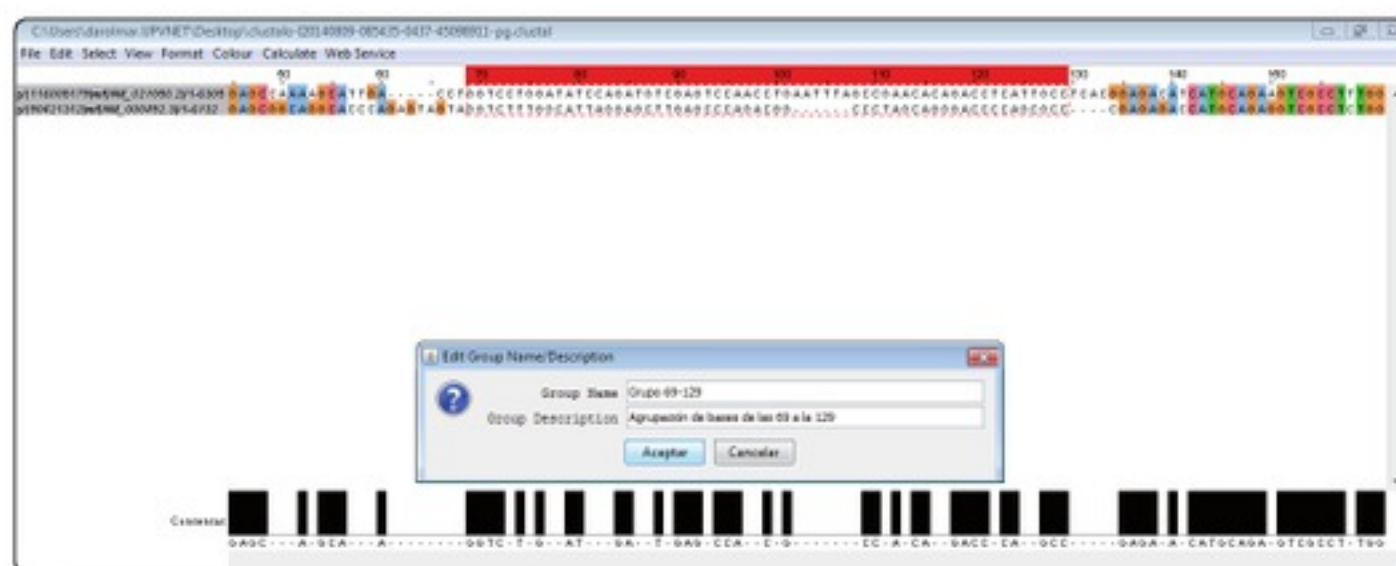


Figura 9.60. Nombre y descripción de un grupo

9.4.2 Reordenación del alineamiento

La reordenación del alineamiento con Jalview es muy sencilla: únicamente hay que seleccionar las secuencias que se desea reordenar y pulsar las flechas del cursor a derecha o izquierda tantas posiciones como se desee. Automáticamente, los nucleótidos (o aminoácidos) de las secuencias seleccionadas se desplazarán en la dirección indicada.

Si seleccionamos la secuencia del ARN del ratón (*accession* NM_021050.2) y la desplazamos 6 posiciones a la derecha, obtendremos la situación de la Figura 9.61:

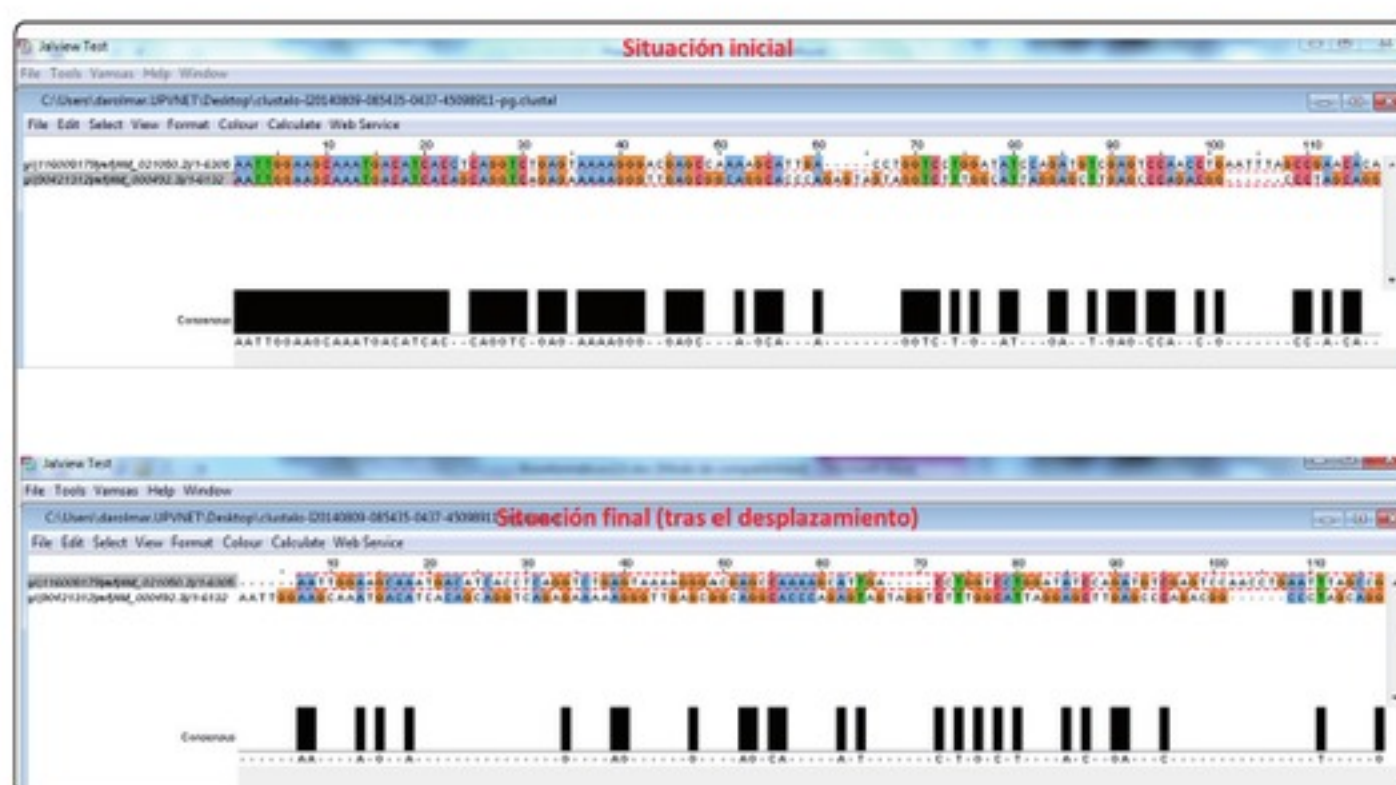


Figura 9.61. Alineamiento modificado por desplazamiento de la primera secuencia



Si se quiere desplazar una secuencia respecto de otras varias, suele ser más sencillo crear un grupo con estas últimas y desplazar el grupo en lugar de la primera secuencia.

9.4.3 Adición y borrado de huecos

Para añadir huecos, únicamente hay que mantener pulsado el botón **Ctrl** y mover el puntero del ratón a voluntad. Inmediatamente, el alineamiento y el consenso se actualizan para reflejar los cambios. Si comparamos la Figura 9.61 con la Figura 9.62, se observan claramente los efectos del desplazamiento de las secuencias agrupadas.

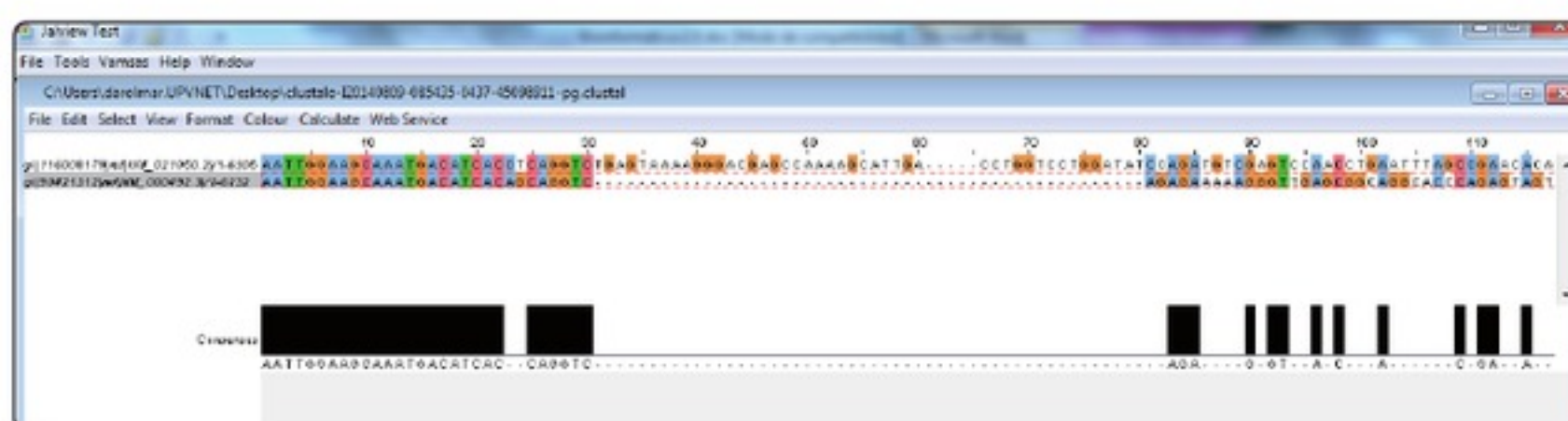


Figura 9.62. Inserción de huecos en el alineamiento

Otra operación habitual en la edición de alineamientos es el borrado de huecos. En Jalview estas operaciones se encuentran agrupadas en el menú **Edición** de la ventana del alineamiento. Por ejemplo, para borrar todos los huecos, dentro de este menú, pinchamos en **Remove All Gaps**:



Figura 9.63. Borrado de huecos

Con lo que obtendremos el alineamiento de la Figura 9.64:



Figura 9.64. Alineamiento tras el borrado de huecos



En estos casos también es buena idea utilizar los grupos, si se desea introducir huecos en las mismas posiciones de secuencias distintas.

9.5 BÚSQUEDA DE SECUENCIAS HOMÓLOGAS CON SIB-BLAST

SIB-BLAST es una herramienta que, dada una secuencia, proporciona un conjunto de secuencias potencialmente homólogas a la secuencia bajo estudio.

En primer lugar, abriremos un navegador e introduciremos la URL de la página principal de SIB-BLAST: <http://bioserv.mps.ohio-state.edu/sib-blast/>. Aparecerá entonces la pantalla de la Figura 9.65:

Figura 9.65. Página principal de SIB-BLAST

SIB-BLAST requiere tres parámetros de entrada. En primer lugar, la secuencia bajo estudio, que puede proporcionarse pegándola en formato FASTA en el área de texto acondicionada para ello o bien subiendo el fichero FASTA con la propia secuencia.

```
>Anonimal
CAGGCAGCCCCACACCCTCCGCCTCCTGCACCGAGAGACATGGAATAAAGCCCCCTGAACCAGCCC
TGCTGTGCCGTCTGTGTGTCTTGGGGGCCCTGGGCCAAGCCCCACTTCCCGGCACTGTTGTGAGC
CCCTCCCAGCTCTCTCCATGCTCTCTGGGTGCCACAGGTGCCAACGCCAGCCAGGCCAGGCCAGCATG
CAGTGGCTCTCCCCAAAGCGGCCATGCCTGTCGGCTGCCTGCTACCCCCACCCTGTGGCTCAGGG
TCCAGTATGGGAGCTGCGGGGGTCTCTGAGGGGCCAGGGGTGGTGGGGCCACTGAGAAATGACTT
CTTGTTTCAGTAGCTCTGGACTCTTGAGGTCCCCAGAGACCTTGTTTCAGGAAAGGGAATGAGAACA
TTCCAGCAATTTTCCCCCACCTAGCCCTCCCAGGTTCTATTTTATAGATTTATTTCTGATGGAGT
CCCTGTGGAGGGAGGAGGCTGGGCTGAGGGAGGGGGTCTGCAGGGCGGGGGGCTGGGAAGGTGG
GGAGAGGCTGCTGAGAGCCACCCGCTATCCCCAGCTCTGGGCAGCCCTGGGACAGTCACACACCC
TGGCCTCGCGGCCCAAGCTGGCAGCCGTCTGCAGCCACAGCTTATGCCAGCCCAGGTCCAGCCAG
ACACCTGAGGGACCCACTGGTGCCTTGGAGGAAGCAGGAGAGGTCAGATGGCACCATGAGCTGGG
GCAGGTGCAGGGACCGTGGCAGCACCGGG
```

Figura 9.66. Secuencia bajo estudio

Adicionalmente, es necesario proporcionar el número de iteraciones del algoritmo de búsqueda que se desea que SIB-BLAST efectúe.



No se recomiendan más de 5 o 6 iteraciones.

Para finalizar, se establece el número máximo de resultados. Hecho esto, pulsando el botón de **Submit** se obtienen los resultados (ver Figura 9.67):

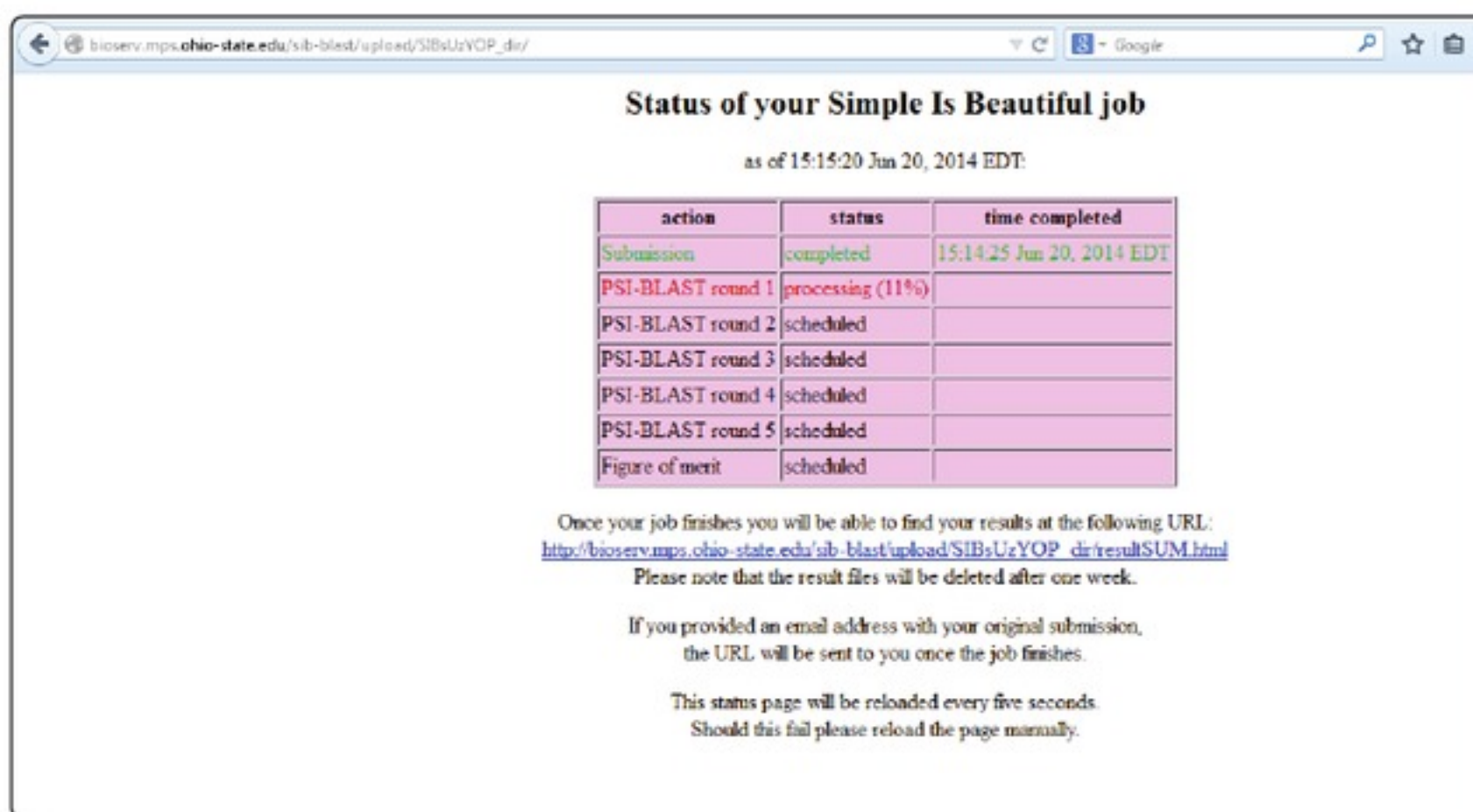


Figura 9.67. Resultados de SIB-BLAST

Si se especifica la dirección de correo electrónico, no es necesario esperar a que acabe el procesamiento del trabajo, sino que cuando este esté listo, llegará una notificación al respecto (Figura 9.68):

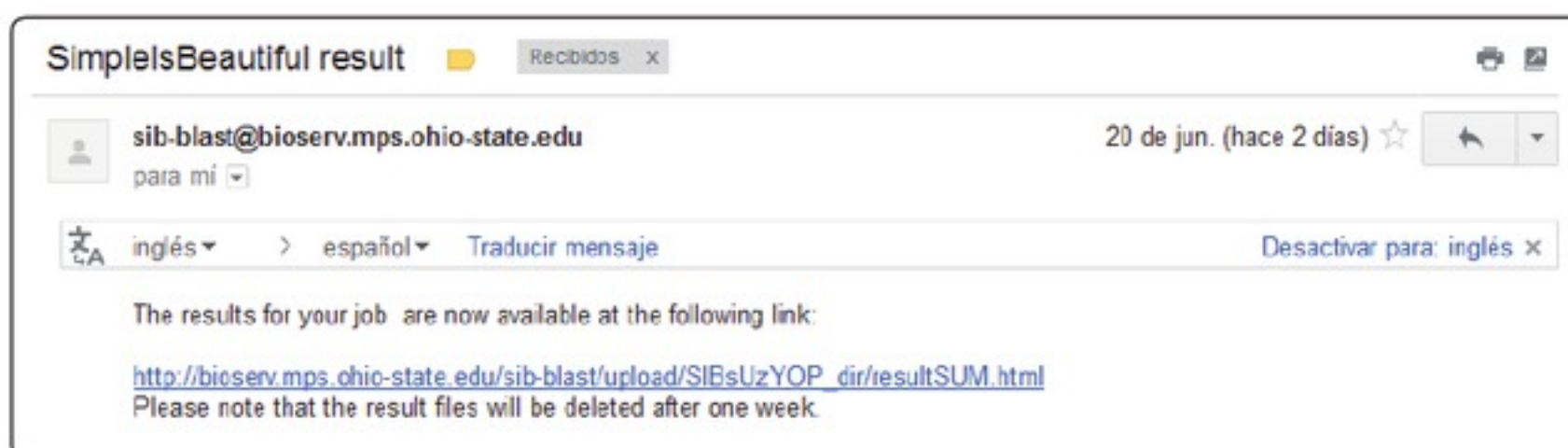


Figura 9.68. Notificación de disponibilidad de resultados

Los resultados se ordenan en una tabla de secuencias putativas homólogas de la secuencia bajo estudio encontradas en la base de datos de proteínas (ver Figura 9.69). Junto con los identificadores de las secuencias, encontramos los valores E en la segunda y la última interacción y la figura de mérito, que se emplea para ordenar la coincidencia de las coincidencias en la base de datos.

9.6 ALINEAMIENTO MÚLTIPLE

En esta práctica trabajaremos con las secuencias de la dehidroascorbato reductasa de varias especies de plantas:

```
>gi|145358113|ref|NM_121676.3| Arabidopsis thaliana DHAR3
(DEHYDROASCORBATE REDUCTASE 1); glutathione dehydrogenase
(ascorbate) (DHAR3) mRNA, complete cds
GGAAATTAAACCACTGCGAAAAATATGATAAGCCTTAGGTTTCAACCAAGCACCACCGCCGGTGT
TTTATCGGCGTCAGTGAGCCGCGCCGGTTTTATCAAGCGGTGCGGTTTCGACTAAACCGGGAAGGG
TTGGCCGGTTTGTACAATGGCGACGGCGGCGAGTCCTCTTGAAATCTGCGTTAAAGCTTCTATC
ACCACTCCCAACAAGCTCGGAGACTGCCCTTTTTGCCAAAAGGTGTTACTGACAATGGAGGAGAA
GAATGTTCTTATGACATGAAAATGGTGGATTTGAGTAACAAACCAGAATGGTTCTTGAAGATTA
GTCCAGAAGGTAAAGTCCCAGTTGTGAAGTTTGATGAGAAATGGGTTCGGATTTCAGATGTTATA
ACACAGGCTTTAGAAGAGAAGTATCCTGAGCCTCCTCTTGCTACCCACCTGAAAAGGCTTCAGT
TGGATCGAAGATCTTTTCCACATTTGTCGGTTTTCTGAAGAGCAAAGATTCAGGAGATGGAAGTG
AGCAAGTTTTGTTGGATGAGCTTACTACATTCAACGATTATATCAAGGACAATGGCCCTTTCATA
AATGGAGAGAAGATCTCGGCAGCAGATTTGTCCTTGGCACCAAAGTTATACCACATGAAGATTGC
ATTGGGACATTACAAGAACTGGTCTGTTCCAGATTCACTTCCTTTCGTCAAATCCTACATGGAGA
ATGTTTTCTCGAGGGAATCATTACGAACACACGGGCGGAAACAGAGGATGTAATTGCTGGTTGG
AGACCAAAGGTTATGGGTAAAGAGATTTTGATGTTGACCTCAAGCTTCCTTGTCTTCTACTAAAT
GTAAAAACATAATAATCAAATCTTCTTCACTGTTACCAATGATATATACCTTGAATCCTTGAT
TGCTTATATGATGAAAACCTATATTGTGTCATAGTGGGGAAGCTTGC
>gi|145335907|ref|NM_101814.3| Arabidopsis thaliana DHAR1
(DEHYDROASCORBATE REDUCTASE); glutathione dehydrogenase
(ascorbate) (DHAR1) mRNA, complete cds
AAAAATGGCCCACTGGTGGGTGGAGAATGGTAATAATTAGTTTAAAAGCTAAGCCTTCTCACTG
ATTAACCTCAATCATTATCCGTCTCTGCAAACAAAGGAAGAAGAATCAAGATGGCTCTGGAAATC
TGTGTGAAAGCTGCTGTTGGTGCTCCTGATCATCTCGGCGACTGTCCGTTTCAGCCAACGGGCTCT
TCTCACACTCGAGGAGAAGAGTCTTACCTACAAAATCCATCTGATTAACCTCTCTGACAAACCCC
AGTGGTTCTTGGACATTAGTCCTCAAGGGAAAGTACCAGTGCTTAAGATCGACGACAAGTGGGTG
ACTGATTCCGACGTCATCGTTGGTATACTCGAGGAGAAGTATCCTGATCCACCACTCAAGACTCC
TGCTGAATTTGCCTCTGTTGGATCCAACATTTTTGGTACTTTTGGGACATTCTTGAAGAGCAAAG
ACTCCAATGACGGATCTGAACATGCCTTGCTTGTTGAGCTAGAAGCTTTGGAAAACCATCTTAAG
AGTCACGATGGCCCTTTTATCGCCGGAGAAAGAGTTTCCGCAGTGGATCTAAGCTTAGCACCAAA
GCTTTACCACCTTCAAGTTGCTCTTGGCCATTTCAAAGCTGGTCTGTCCCTGAGAGCTTTCCCC
ATGTCCATAACTACATGAAGACTCTGTTCTCGCTCGACTCTTTTGAGAAAATAAGACCGAGGAA
AAGTATGTGATCTCTGGATGGGCTCCCAAGGTAAACCCTTGAACTATGCACCGTTATGAGATCG
TTGGTGTGGTAATGTTGTTCAAGGTCTCTCTTATATCAGTCAAATAATGTACTTGGACCTTTT
TATGTAATGTACTGTATCAATCAAATAAGAAGCCTATTTTGAAATAAGATTTGCCTCCATATC
>gi|123187086|gb|EF195334.1| Solanum tuberosum dehydroascorbate
reductase (DHAR2) mRNA, complete cds
AACATGTCGACCGCAAAGATAACACCATCAGCTGCTTCATTTGCGACTTCTATCAAACACCTTGC
GGGCATTCAACTACCTCGACTCCAAAACACCATTTATACCTCCAATTCCACTAAGTTTAGAGCAC
```



```
CCAGAAGAGCTTTTACTGTATCAATGGCGGCTTCACTCGACACCCCTCTTGAAGTTTGCGTCAAA
CAATCAATTACAACCTCCTAACAAGCTCGGCGACTGCCCATTCACTCAGAGGGTTTTGCTTACGTT
GGAGGAAAAGCACCTTCCATATGACATGAAGTTTGTTGATTAAAGTAACAAGCCTGACTGGTTTT
TGAAGATAAGCCCCGAAGGTAAAGTTCCACTTATTAAGCTTGACGAGAAATGGGTTCAGATTCA
GATGTCATCACACAGGCACTGGAGGAGAAGTTCCCTGAACCTCCGCTGACAACTCCTCCTGAGAA
GGCTTCCATTGGATCAAAGATCTTCCCGAAGTTTGTTGCTTTTCTGAAAAGCAAAGACCCCACTG
ATGGAACAGAGCAGGCTTTACTTGATGAGCTGACAGCTTTCAATGATTACCTTAAAGAAAATGGT
CCATTTATCAACGGAAATGAGGTATCTGCTGCTGATTTGTCGCTTGGACCAAAGCTATATCATTT
AGAAATATCTTTGGGGCACTATAAGAATTGGTCTATTCCAGATTCACTTTCCTACGTGAAATCAT
ACATGGAGAGTACATTCTCCAGGGAATCATTCAACACGCGGGCACTAAAAGAGGACGTCATT
GAAGGTTGGCGACCAAAGTCATGGGTTAGACAAACTATATCATCTTTTGCATTTCTGAGGATTA
GATTTTTGTGACAAGGTATAGTAAGCTAGCATTTGGAAGGCTGTATGACAGTTCTTTGCCATGTA
TATTGTTATTAAAACATATACTCATCCTTGTTTGACTGAATGATAGCCTGA
```

>gi|66475037|gb|AY971874.1| *Lycopersicon esculentum*

dehydroascorbate reductase (DHAR2) mRNA, complete cds

```
CAGGCAACATGTCGACCGCAAAGATAACACCGTCAGCTGCTTCATTTGCGACTTCTATCAAACAC
CTTGCGGGCATTCAACTACCTCGACGCCAAAGCACCATTTTTACCTCCAATTCCACGAAGTTCAG
AGCACCCAGAAGAGGTTTTACTGTATCAATGGCGGCTTCAATCGAAACCCCTCTTGAAGTATGCG
TCAAACAATCAATTACAACCTCCTAACAAGCTCGGTGACTGCCCATTCAACCAGAGGGTTTTGCTT
ACGTTGGAGGAGAAACACCTTCCATATGACATGAAGTTTGTTGACTTGAGTAACAAGCCTGACTG
GTTTTTGAAGATAAGCCCTGAAGGTAAAGTTCCTCTTATTAAGCTTGACGAGAAATGGGTGCCAG
ATTCAGATGTCATCTCACAGGCACTGGAGGAGAAGTTCCCCAAACCTCCGCTGACAACTCCTCCT
GAGAAGGCTTCCGTCGGATCAAAGATTTTCCCCAAGTTTGTTGCTTTCCTGAAAAGCAAAGACTC
CGGTGATGGAACAGAGCAGGCTTTACTTGATGAGCTGACAGCTTTCAATGATTACCTTAAAGAAA
ATGGTCCATTTATCAACGGAAATGAGGTATCTGCTGCTGATTTGTCGCTTGGACCAAAGCTATAT
CATTTAGAAATAGCTTTGGGGAACTATAAGAATTGGTCTATTCCAGATTCACTTTCCTACATGAA
ATCATACATGAAGAGTATATTCTCCAGGGAATCATTCAATCACACGCGGGCACTAAAAGAGGATG
TCATTGAGGGTTGGCGACCAAAGTCATGGGTTAGACAAACTATATCATCTTTTGCATTTCTGAG
GATTAGGTTTTTGTGCAAGGTATAGTAAGCTAGCATTTGGAAGGCTGTAAGACAGTTCTTTGCC
CTGTATATTGTTATTAAAACACATACTCATCCTTGTTAGACTGAATGATAGCCTGAGTTATATAT
GTAATACATACTTCCTAAGGCTTGTGCATAAAAAAAAAAAAAAAAAAAAAA
```

>gi|28192426|gb|AY074787.1| *Nicotiana tabacum* dehydroascorbate
reductase (DHAR) mRNA, complete cds

```
CGTTGCTGTCGGTTCCAAAAGAAAGACACTATATTCACCCAGAGTAGAATTCTCTAGTCCTTACC
AACTACAAAGCGAAAAAGCTGTAGATCAATGGCTGTTGAAATCTGTGTCAAGGCTGCTGTGGGT
GCCCTAATGTCCTCGGAGACTGTCCATTTAGCCAAAGGGCACTTCTGACATTGGAGGAAAAGAA
AGTGCCTTACAAGATGCACTTGATCAATGTTAGTGACAAGCCCAAATGGTTCTTGGAAGTGAACC
CAGAAGGAAAAGTTCCAGTGATCAAGTTTGATGAAAAATGGATCCCTGATTCTGATGTTATTGTT
GGGCTTCTTGAAGAGAAATACCCAAATCCCTCTCTCTCTAGTCCCCCTGAATTTGCTTCTGTGGG
CTCGAAAATATTTCTTCTTTGTCTCATTTTCGGAAGAGCAAGGATGCTAGTGACGGTACTGAGC
AGGCTCTGCTCGACGAGTTAAAGGCTTTGGAAGAGCATCTCAAGGCTCACGGACCATATGTCAAT
GGGGCGAATATTTGTTCAAGTCGATTTGAGTTTGCTCCGAAACTGTACCATCTTGAGGTGGCTCT
TGGCCATTTCAAGAAGTGGAGTGTACCTGAAAGCTTGAGTCATGTGCGTAAATACATGAAGTTGC
TCTTCGAGCGAGAGTCTTCCAGAAAACCAAGGCTGCAAAAGAGTATGTCATTGCAGGATGGGCT
CCAAAGGTCAATCCATGAACCGATTCAATAATTATAATCCCGTTGTTTCGCAGGAAGTTAGCAGTT
```



```
GAGGATACAGCATTTTGAAATATGAATGTATCTCGTAAGATCTAAAAATTGTTAAATGTTGGATC
ATGCTTGTACTGCTCTTTATGTTCTAATAAATAAGTCATGTTCTAAAAAAAAAAAAAAAAAAAAAA
AA
```

Figura 9.71. Secuencias bajo estudio



Gracias al alineamiento múltiple es posible realizar diversos análisis: filogenia, búsqueda de motivos funcionales, etc. De ahí la importancia de dominar el manejo de estas herramientas.

9.6.1 Alineamiento múltiple con Clustal Omega

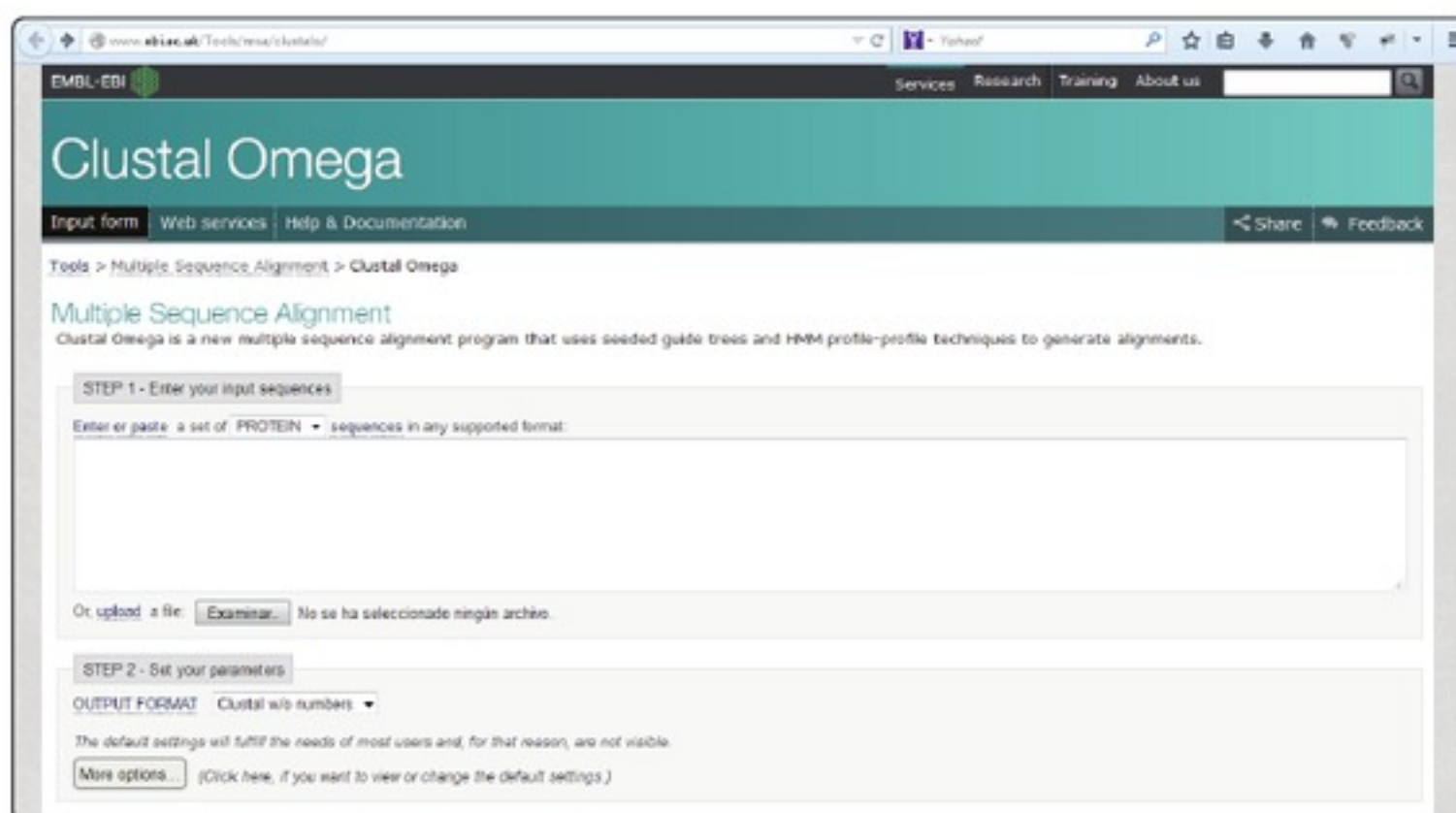
Clustal Omega es un programa de alineamiento múltiple de secuencias de ADN o de proteínas. Para realizar los alineamientos, ClustalO sigue una estrategia progresiva en tres pasos:

- Alineamiento global de pares: lanza un alineamiento de pares entre cada una de las secuencias y el resto. Es decir, para N secuencias, se efectuarán:

$$N_{\text{alineamientos}} = \frac{n!}{2(n-2)!}$$

- Construcción del árbol guía: a partir de la puntuación de cada alineamiento de pares, se construye un árbol que describe la relación entre las secuencias.
- Alineamiento múltiple: siguiendo el árbol guía, se comienzan los alineamientos múltiples.

Para utilizar Clustal Omega, en primer lugar, accederemos a su página principal (<http://www.ebi.ac.uk/Tools/msa/clustalo/>). Aparecerá la pantalla principal de Clustal Omega (ver Figura 9.72), en la que se indican los pasos que hay que seguir para utilizar la herramienta.



EMBL-EBI Clustal Omega

Services Research Training About us

Input form Web services Help & Documentation

Tools > Multiple Sequence Alignment > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments.

STEP 1 - Enter your input sequences

Enter or paste a set of **PROTEIN** sequences in any supported format:

Or upload a file: No se ha seleccionado ningún archivo.

STEP 2 - Set your parameters

OUTPUT FORMAT: Clustal w/o numbers

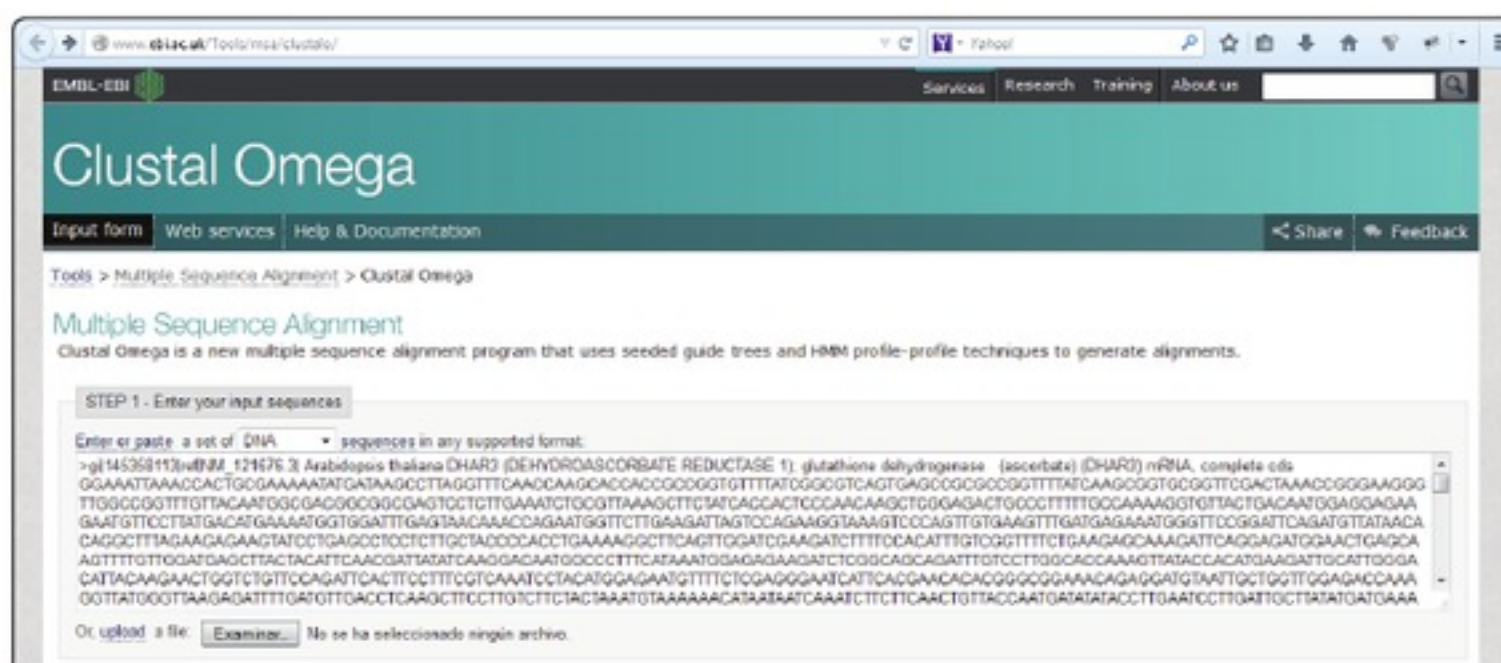
The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)

Figura 9.72. Página principal de Clustal Omega

A continuación, pegamos las secuencias que vamos a alinear y seleccionamos la opción **DNA** en lugar de **Protein** (opción por defecto). Los datos necesarios para realizar un análisis con Clustal Omega se agrupan en cuatro secciones de la página principal: datos de las secuencias, opciones del alineamiento, parámetros del algoritmo y configuración del envío.

En la sección de secuencias, como viene siendo habitual, es posible pegar las secuencias en el área de texto o bien adjuntarlas en un fichero en formato FASTA e indicar si se trata de una secuencia de nucleótidos o de aminoácidos, aunque la propia aplicación nos advierte del error, si realizamos la selección incorrectamente.



EMBL-EBI Clustal Omega

Services Research Training About us

Input form Web services Help & Documentation

Tools > Multiple Sequence Alignment > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments.

STEP 1 - Enter your input sequences

Enter or paste a set of **DNA** sequences in any supported format:

```
>gl145358113.m121676.3 Arabidopsis thaliana (DHAR) (DEHYDROASCORBATE REDUCTASE 1): glutathione dehydrogenase (ascorbate) (DHAR) mRNA, complete cds
GGAAATTAACCACTGGGAAATATGATAAGCCTTAGGTTTCAACCAAGCACCACCGGGGTTTATCGGGCTCAGTGAGCGCGCGGGTTTATCAAGCGGTGCGGTTGACTAAACCGGGAAGGG
TTGGCCGGTTTGTACATGGCGAGCGCGCGGCTCTTGAATCTGGTTAAAGCTTCATCAGCACTCCCAACAGCTGGAGACTGGCGCTTTTGGCAAAAGGTGTACTGACATGGAGGAGAA
GAATGCTTCTATGACATGAAATGGTGGATTGAGTAACAAACAGAAATGGTTCTTGAAGATTAGTCCAGAGGTAAGTCCAGTTGTGAAGTTTGATGAGAAATGGGTTCCGGATTCAGATGTATAACA
CAGGCTTAGAAGAGAGATATCTCAGGCTCTCTTGTCTAGCCACCTGAAAGGCTTCAGTTGGATCGAAGATCTTTTCACATTGTGCGTTTCTGAAGAGCAAGATTCCAGAGATCGAATGAGCA
AGTTTGTGGATGAGCTTACTACATTCACGATTATCAAGGAGAAATGGCGCTTTCATAAATGAGAGAGAGATCTGGGCAAGCAGATTTGCTTGGCAGCAAGGTTATACCACATGAGATGCAATGGGA
CATACAGAACTGGTCTGTTCCAGATTCAGTTCTTCTGCTCAATCTACATGGAGAAATGTTTCTGGAGGGAAATCATTCACGACACACCGGCGGAAACAGAGGATGTAATGCTGCTGGAGACCAA
GGTTATGGGTTAAGAGATTTGATGTTGACCTCAAGCTCTCTGCTTCTACTAAATGTAAAAACATAAATCAATCTTCTCAACTGTTACCAATGATATACCTTGAATCTTGTATGATGATGAAA
```

Or upload a file: No se ha seleccionado ningún archivo.

Figura 9.73. Sección de secuencias

En la sección de opciones del alineamiento (ver Figura 9.74), podemos seleccionar las características del alineamiento.

Figura 9.74. Parámetros de configuración del alineamiento

En primer lugar, indicamos si queremos desalinear las secuencias de entrada (**Dealign input sequences**) eliminando de las mismas los huecos.

A continuación, escogemos el formato del fichero de alineamiento entre alguno de los valores recogidos en la Tabla 9.1:

Formato	Descripción	Selección
Clustal sin números	Formato Clustal sin la numeración de las bases/residuos incluida	Clustal
Clustal con números	Formato Clustal con la numeración de las bases/residuos incluida	clustal_num
Pearson/FASTA	Formato Pearson o FASTA	fa
MSF	Formato MSF (<i>Multiple Sequence File</i>)	Msf
PHYLIP	Formato de alineamiento entrelazado PHYLIP	Phylip
SELEX	Formato de alineamiento SELEX	Selex
STOCKHOLM	Formato de alineamiento Stockholm	Stockholm
VIENNA	Formato de alineamiento Vienna	Vienna

Tabla 9.1. Formatos del fichero de alineamiento soportados por Clustal0

Especialmente cuando el número de secuencias es elevado, resulta adecuado habilitar la generación rápida del árbol guía, mediante el muestreo de las secuencias de entrada y su representación como vectores. Esta opción se controla con el parámetro **mBed-like Clustering Guide Tree**. Si, además, se desea que se utilice la misma filosofía en las sucesivas iteraciones, hay que establecer el parámetro **mBed-like Clustering Iteration** al valor **true**.

Finalmente, se especifican los parámetros relacionados con las iteraciones del algoritmo, que son:

- Número de iteraciones combinadas (**Number of Combined Iterations**): número de iteraciones del algoritmo de alineamiento, entre las realizadas sobre el árbol guía y las HMM (*Hidden Markov Model*).
- Número máximo de iteraciones del árbol guía (**Max Guide Tree Iterations**): una vez establecido el número de iteraciones combinadas, este parámetro se puede cambiar para limitar el número de iteraciones del árbol guía, dentro de los límites de las iteraciones combinadas árbol guía/HMM.
- Número máximo de iteraciones HMM: similar al anterior, pero controlando las iteraciones del modelo de Markov.

Por último, se especifica si el orden de las secuencias en el alineamiento (**order**) será el mismo en el que se proporcionaron las secuencias de entrada (**input**) o si, por el contrario, será el propio alineamiento el que determine el orden (*alignment*).

Finalmente, se puede introducir una dirección de correo electrónico para que Clustal Omega le notifique que el análisis ha concluido. Una vez que ClustalO ha recibido los datos de su análisis, se le mostrará una pantalla similar a la de la Figura 9.75:

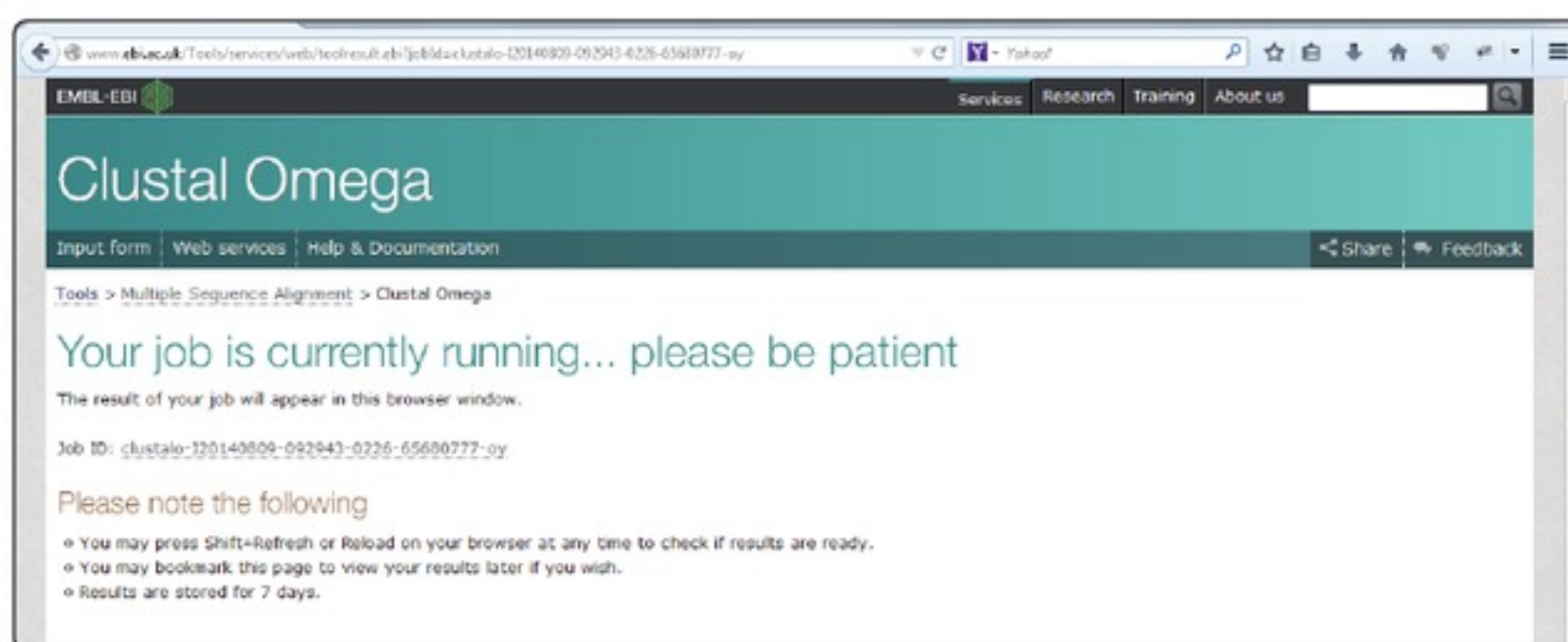


Figura 9.75. Trabajo procesándose

Si se ha escogido la notificación por correo electrónico, cuando el informe esté listo llegará una notificación (ver Figura 9.76):

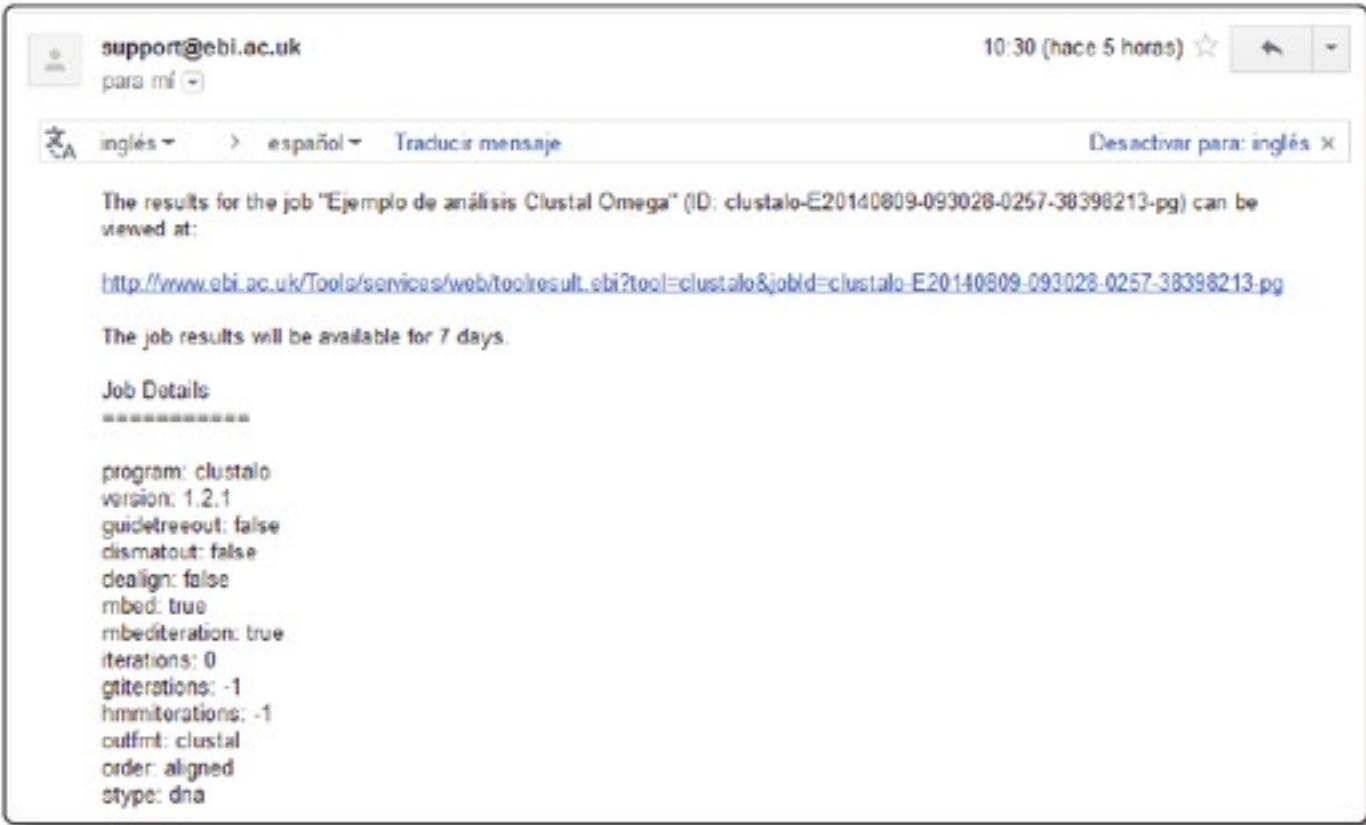


Figura 9.76. Correo informativo con los resultados del trabajo

Al pinchar en el enlace, se abre una ventana del navegador (ver Figura 9.77) con los resultados del análisis.

En la primera pestaña, **Alignments**, podemos descargar el fichero de alineamiento, colorear los resultados para facilitar su análisis visual o enviarlo a una herramienta de filogenia.

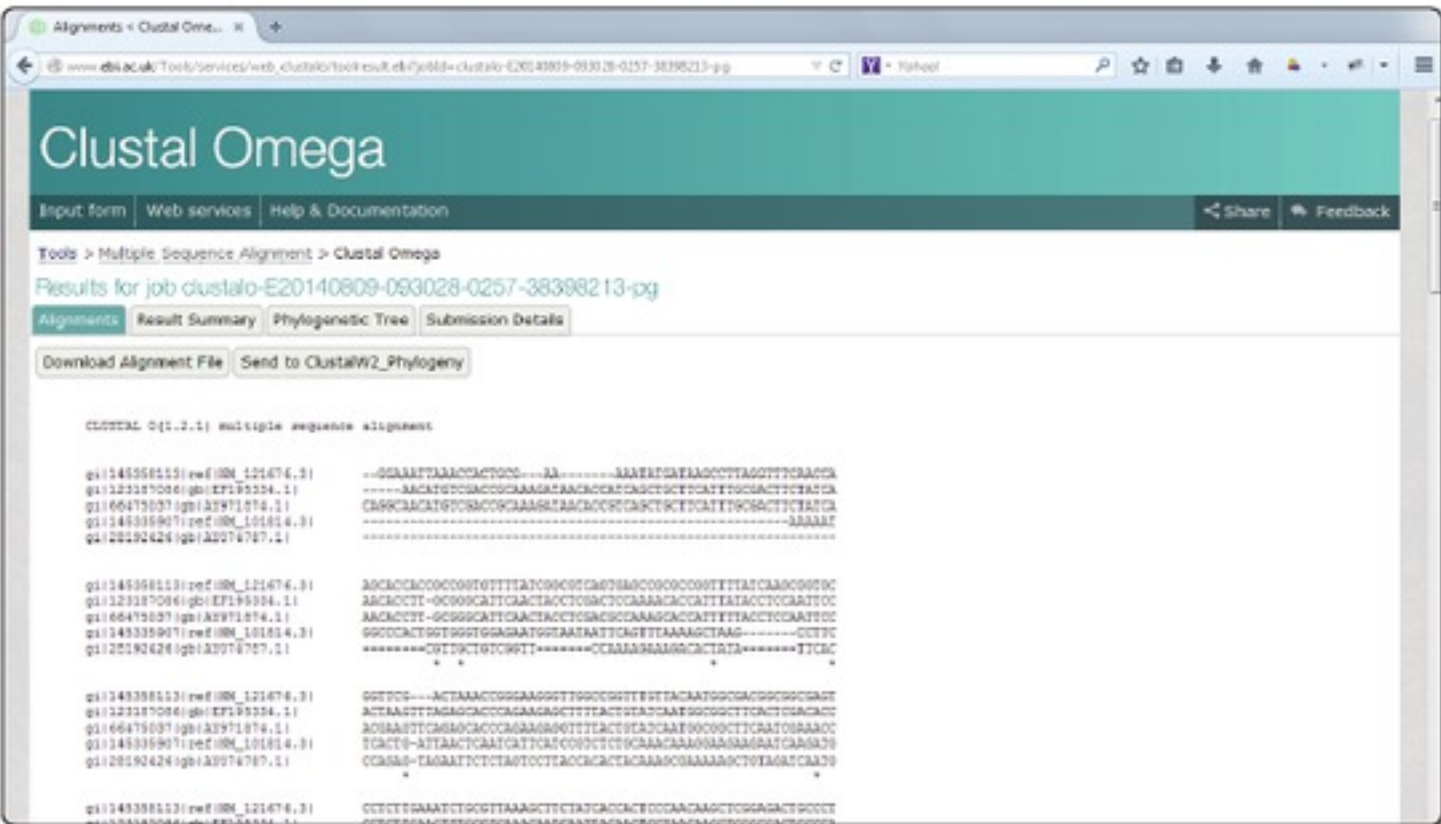


Figura 9.77. Sección de alineamientos

Si examinamos más detenidamente uno de los alineamientos (ver Figura 9.78), veremos que se utilizan algunos símbolos justo debajo de las líneas correspondientes a los fragmentos de las secuencias. El carácter '.' indica que hay una coincidencia en las secuencias que se alinean. Si, por el contrario, el carácter empleado es el ':', entonces la coincidencia es alta. Finalmente, si encontramos el carácter '*', la coincidencia es exacta.

```

CLUSTAL O(1.2.1) multiple sequence alignment

gi|145358113|ref|NM_121676.3|      --GGAATTAAACCACTGCG---AA-----AATATGATAAGCCTTAGGTTTCAACCA
gi|123187086|gb|EF195334.1|      -----AACATGTCGACCGCAAAGATAACACCATCAGCTGCTTCATTGCGACTTCTATCA
gi|66475037|gb|AY971874.1|      CAGGCAACATGTCGACCGCAAAGATAACACCGTCAGCTGCTTCATTGCGACTTCTATCA
gi|145335907|ref|NM_101814.3|      -----AAAAAT
gi|28192426|gb|AY074787.1|      -----

gi|145358113|ref|NM_121676.3|      AGCACCAACCGCCGGTGTGTTTATCGGCGTCAGTGAGCCGCGCCGGTGTGTTTATCAAGCGGTGC
gi|123187086|gb|EF195334.1|      AACACCTT-GCGGGCATTCAACTACCTCGACTCCAAACACCATTTATACCTCCAATTCC
gi|66475037|gb|AY971874.1|      AACACCTT-GCGGGCATTCAACTACCTCGACGCCAAGCACCATTTTACCTCCAATTCC
gi|145335907|ref|NM_101814.3|      GGCCCACTGGTGGGTGGAGAAATGGAATAATTGAGTTTAAAGCTAAG-----CCTTC
gi|28192426|gb|AY074787.1|      -----CGTTGCTGTCGGTT-----CCAAAGAAAGACACTATA-----TTCAC
                                   *  *                               *          *

```

Figura 9.78. Interpretación de un alineamiento múltiple

La segunda pestaña (ver Figura 9.79) resume los resultados ofreciendo información general sobre el alineamiento:

Clustal Omega

Input form | Web services | Help & Documentation

Tools > Multiple Sequence Alignment > Clustal Omega

Results for job clustalo-E20140809-093028-0257-38398213-pg

Alignments | **Result Summary** | Phylogenetic Tree | Submission Details

Input Sequences

[clustalo-E20140809-093028-0257-38398213-pg.input](#)

Tool Output

[clustalo-E20140809-093028-0257-38398213-pg.output](#)

Alignment in CLUSTAL format

[clustalo-E20140809-093028-0257-38398213-pg.clustal](#)

Phylogenetic Tree

[clustalo-E20140809-093028-0257-38398213-pg.ph](#)

Percent Identity Matrix

[clustalo-E20140809-093028-0257-38398213-pg.pim](#)

Jalview

[Start Jalview](#)

Figura 9.79. Resumen de los resultados

Desde esta misma pestaña, es posible visualizar el fichero de alineamiento en Jalview (ver Figura 9.80):



Figura 9.80. Visualización del alineamiento en Jalview

Las otras dos pestañas restantes proporcionan información para la construcción del árbol filogenético de las secuencias (**Guide Tree**) y el propio árbol (**Phylogenetic Tree**).

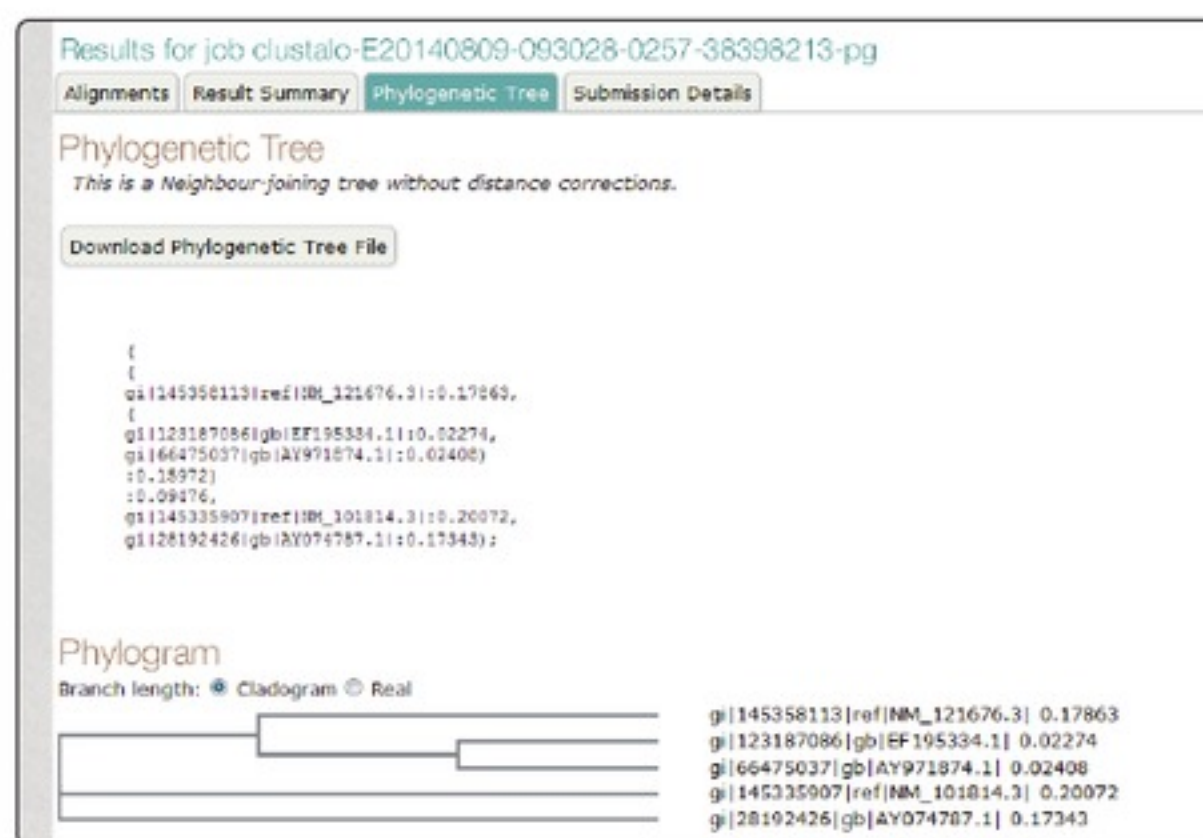


Figura 9.81. Árbol filogenético correspondiente a las secuencias bajo estudio

Finalmente, existe una pestaña adicional en la que se recoge información del envío (**Submission details**).

9.6.2 Alineamiento múltiple con MUSCLE

En primer lugar, abrimos la página principal de MUSCLE (<http://www.ebi.ac.uk/Tools/msa/muscle/>) y establecemos la configuración del alineamiento (ver Figura 9.82):

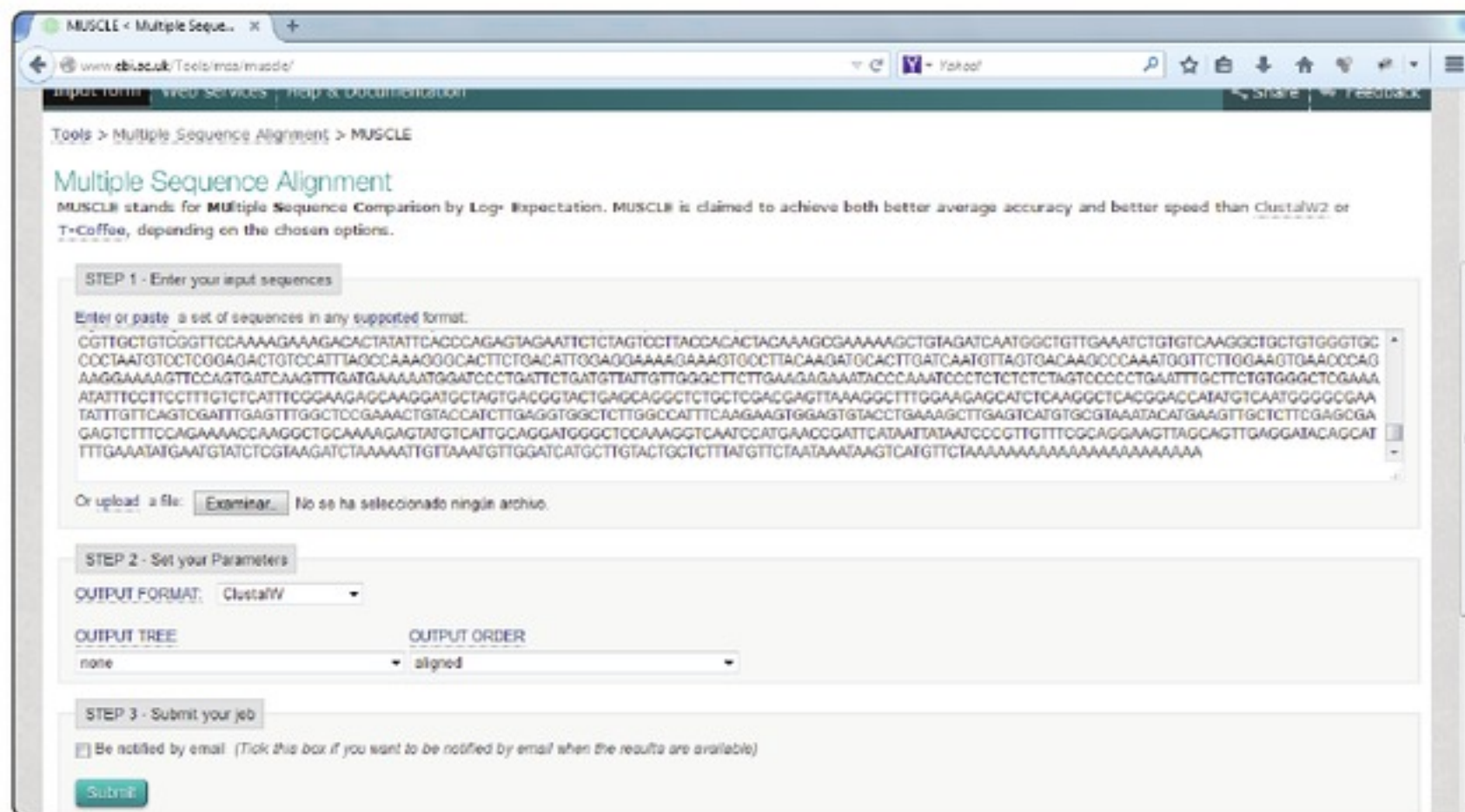


Figura 9.82. Configuración del alineamiento con MUSCLE

En la primera sección, pegamos las secuencias que se quieren alinear o subimos un fichero en formato FASTA con las mismas.

En la segunda sección, introducimos los parámetros de entrada, que son:

- Formato de salida: formato que se dará al alineamiento de salida. Se suele emplear la opción por defecto, ClustalW.
- Árbol de salida: hace referencia al árbol guía que se tomará como base para generar los alineamientos múltiples. El valor por defecto es ninguno (**none**), aunque puede indicarse que se utilice uno para la primera iteración (**tree1**) o para la segunda (**tree2**).
- Orden de salida en que se mostrarán las secuencias en el alineamiento final. Actualmente, únicamente es posible el valor **aligned**, que significa que se mostrarán según el árbol guía.

Si pinchamos en el botón **Submit**, se lanzará el alineamiento múltiple y, al cabo de unos instantes, se mostrará una pantalla de resultados similar a la del epígrafe anterior.

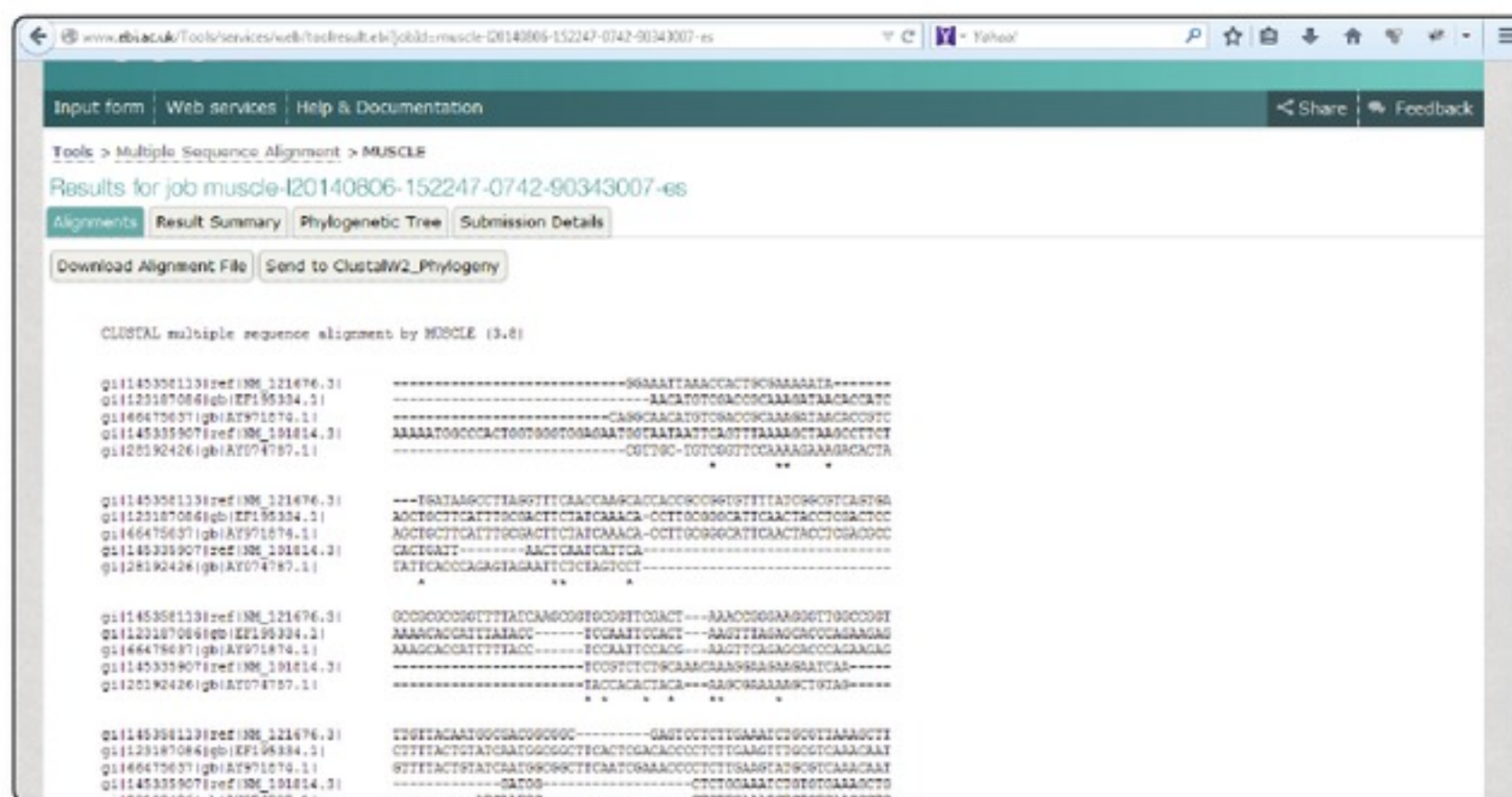


Figura 9.83. Resultados del alineamiento con MUSCLE

Si nos detenemos a observar el residuo que con CLUSTALW detectamos que estaba mal alineado, veremos que MUSCLE lo ha alineado correctamente (ver Figura 9.84):

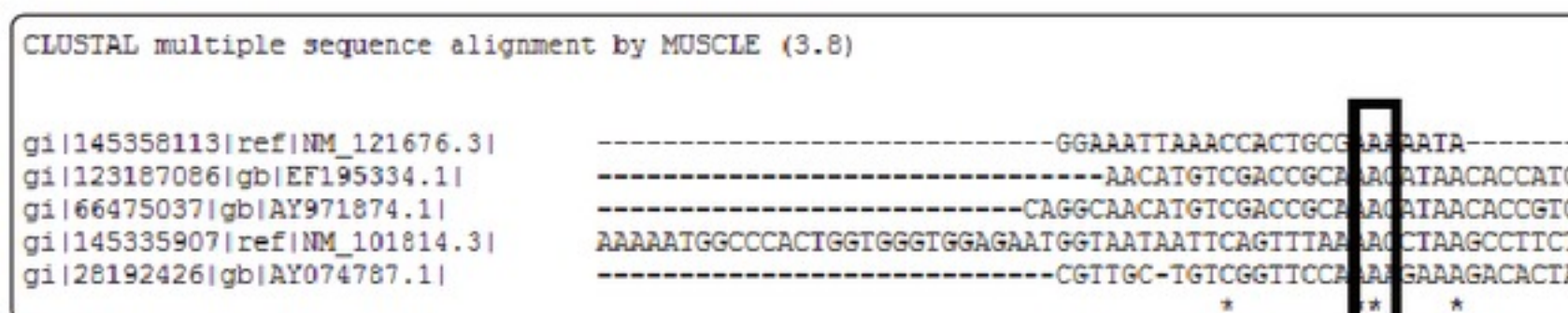


Figura 9.84. Alineamiento con MUSCLE

9.6.3 Alineamiento múltiple con T-Coffee

La herramienta T-Coffee está disponible en <http://www.ebi.ac.uk/Tools/msa/tcoffee/>. La configuración del alineamiento, en este caso, consiste en establecer la matriz de sustitución que se desea emplear al generar el alineamiento múltiple,

escogiendo entre ninguna (**none**), **BLOSUM** o **PAM** (se trata de una PAM 350). También se indica si se desea que las secuencias se ordenen según el alineamiento (**align**) o que se mantenga el orden de entrada (**input**). Una vez establecida la configuración, pulsando el botón **Submit** se obtienen los resultados:

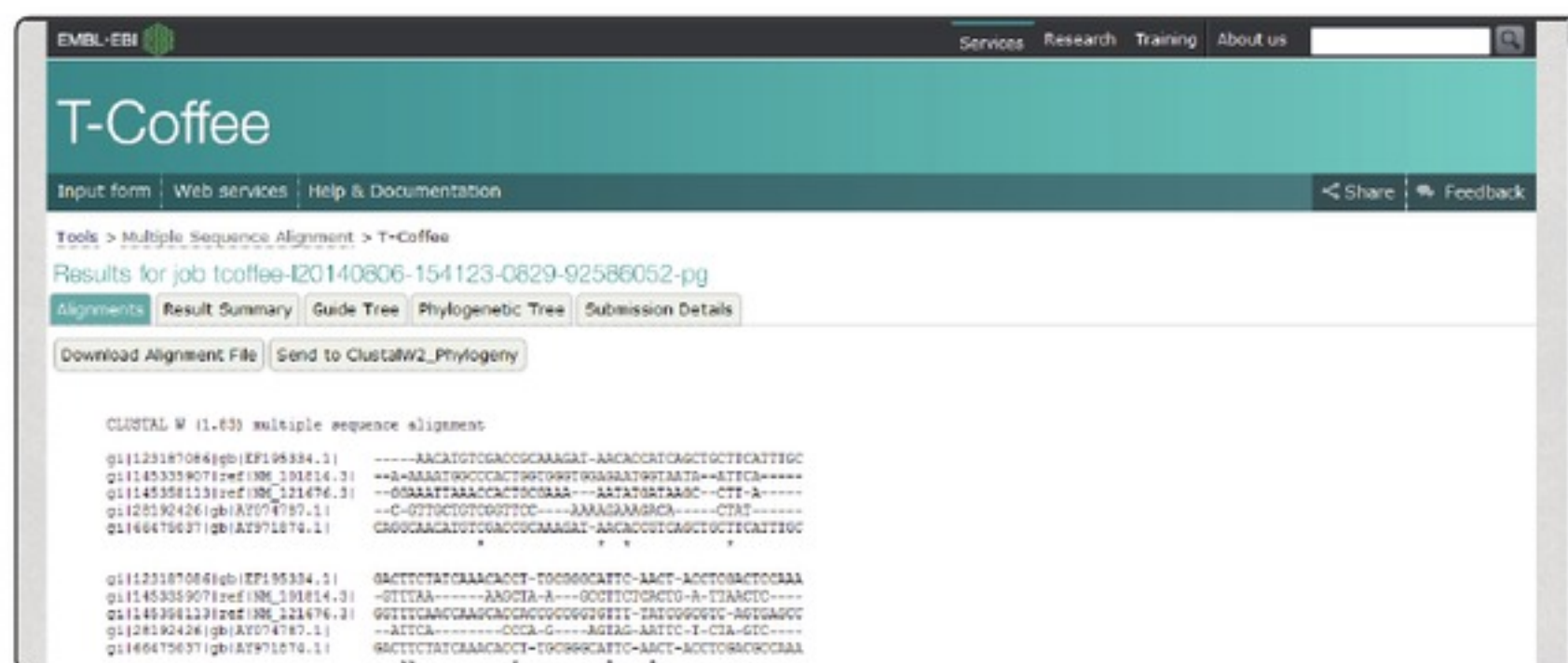


Figura 9.85. Alineamiento con T-Coffee



10

PROTEÓMICA

La Proteómica, en su acepción más simple, es el estudio del conjunto de proteínas de un organismo y consiste en el análisis sistemático de dichas proteínas, su estructura, función y regulación del organismo en cuestión.

Por lo general, el estudio de las proteínas es mucho más complejo que el del ADN ya que no se dispone de técnicas de amplificación con la PCR sino que el estudio de las proteínas debe realizarse *in vivo*. Además, no existe una aproximación estándar para el estudio de las proteínas.

10.1 GENERALIDADES

Las proteínas representan el 50% del peso seco de los seres vivos de la biosfera (ver Tabla 10.1). Se trata de compuestos químicos de gran peso molecular formados por una secuencia determinada de elementos más simples denominados aminoácidos y que desempeñan funciones estructurales (colágeno del tejido conjuntivo), transportadoras (hemoglobina de la sangre), nutritivas, inmunológicas, hormonales y catalíticas. En ellas reside la información funcional de la célula.

Componente	% del peso total
Agua	69
Iones inorgánicos	1
Metabolitos pequeños	3
Proteínas	18
ARN	1,1
ADN	0,25
Fosfolípidos	3
Otros lípidos	2
Polisacáridos	2

Tabla 10.1. Composición aproximada de una célula típica de mamífero

10.2 ESTRUCTURA DE LAS PROTEÍNAS

Conocer la estructura de una proteína permite, no solo conocer su influencia en el comportamiento físico-químico del organismo, sino que sienta las bases para averiguar cómo interferir en ellas e incluso, diseñar proteínas “a medida”.

En esencia, todas las proteínas son polímeros de aminoácidos, caracterizados por presentar simultáneamente un grupo carboxilo ($-\text{COOH}$) y un grupo amino ($-\text{NH}_2$), junto con un radical lateral (R) que es el que diferencia unas de otras.

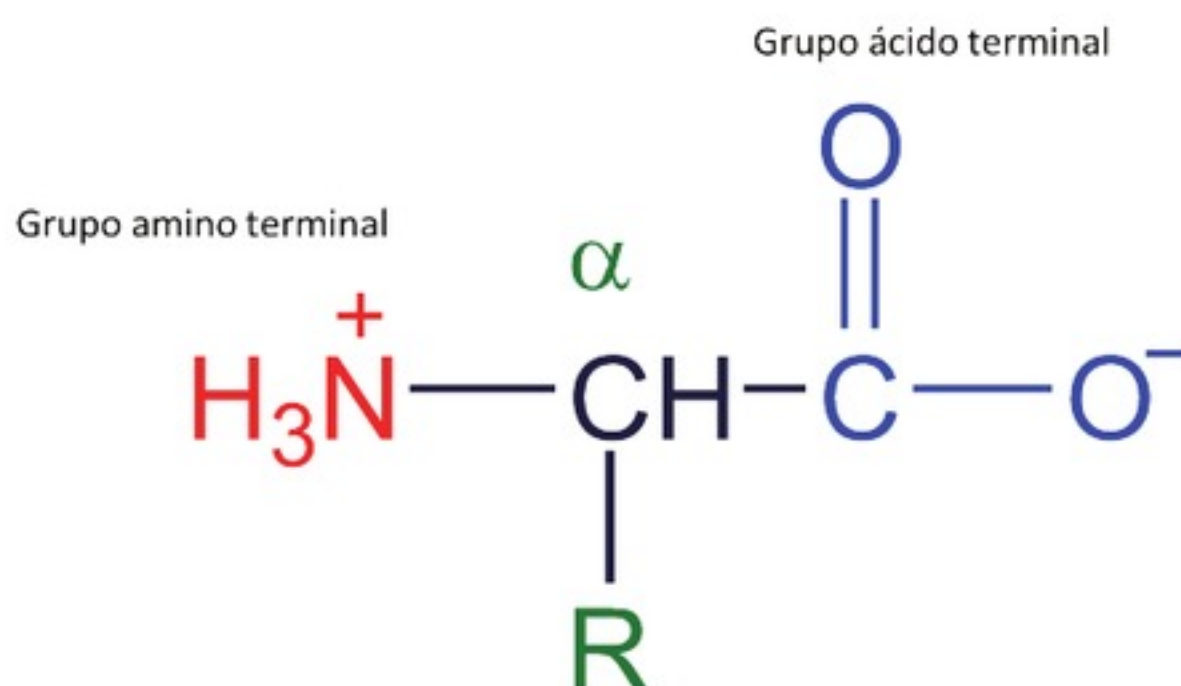


Figura 10.1. Estructura de un aminoácido

En la naturaleza viva, 20 son los aminoácidos más comunes. Los aminoácidos se combinan formando compuestos denominados polipéptidos mediante uniones llamadas enlaces peptídicos. El enlace peptídico se produce cuando el grupo carboxilo (CH_2) de un aminoácido reacciona con el grupo amino del siguiente desprendiendo una molécula de agua.

Al orden en que los aminoácidos se ordenan para formar proteínas se le denomina estructura primaria. Es altamente específica de cada proteína y un cambio de orden o sustitución de un único aminoácido supone la alteración de la molécula.

Esta cadena polipeptídica adopta una disposición espacial denominada estructura secundaria. En esta estructura espacial la cadena principal forma la estructura central y cadenas laterales que se extienden por fuera de la hélice (hélice alfa) o en hoja plegada (plegamiento beta).

La estructura terciaria es la estructura tridimensional de la cadena polipeptídica. Cuando hay más de una cadena polipeptídica, la representación de cómo se interconectan entre sí y de su organización es la estructura cuaternaria.

La información necesaria para la síntesis de proteínas está contenida en el ácido desoxirribonucleico (ADN) que se forma en el material genético del microorganismo.

10.3 MÉTODOS DE PREDICCIÓN

La primera familia de métodos de predicción son los llamados métodos *ab initio*, que parten de la suposición de que, para conocer la estructura de la proteína, toda la información necesaria se encuentra en su secuencia de aminoácidos. Su principal ventaja es que permite modelar proteínas tomando como dato de partida su secuencia, aunque son computacionalmente muy costosos. Todo ello hace que en la práctica se utilicen como apoyo a otras técnicas más potentes.

El segundo conjunto de métodos de predicción es el modelado por homología. En este caso, se trata de buscar proteínas cuyas secuencias tengan cierto grado de similitud (por encima del 30%) y, apoyándose en el mismo, suponer que la estructura de la proteína bajo estudio y la de la proteína encontrada también serán similares.

Sin embargo, cuando la similitud entre la proteína bajo estudio y la referencia es baja, no es posible aplicar el modelado por homología. En estos casos, se utiliza el reconocimiento de pliegues o *threading*, que consiste en someter la secuencia de la proteína bajo estudio a diferentes pliegues conocidos y evaluar cómo coincide dicha proteína en cada uno de ellos.

10.4 MODELADO POR HOMOLOGÍA

El modelado por homología es un proceso iterativo que consiste en repetir las siguientes etapas hasta conseguir una calidad razonable en los resultados (ver Figura 10.2):

- Identificación de estructuras conocidas o referencias relacionadas con la secuencia bajo estudio utilizando métodos BLAST.
- Alineamiento de la secuencia bajo estudio con las referencias (por ejemplo, con CLUSTAL).
- Construcción del modelo de la secuencia bajo estudio partiendo del alineamiento realizado.
- Evaluación del modelo. Si se supera el control de calidad, el modelo se considera bueno. En caso contrario, se efectúa una nueva iteración revisando las desviaciones encontradas.

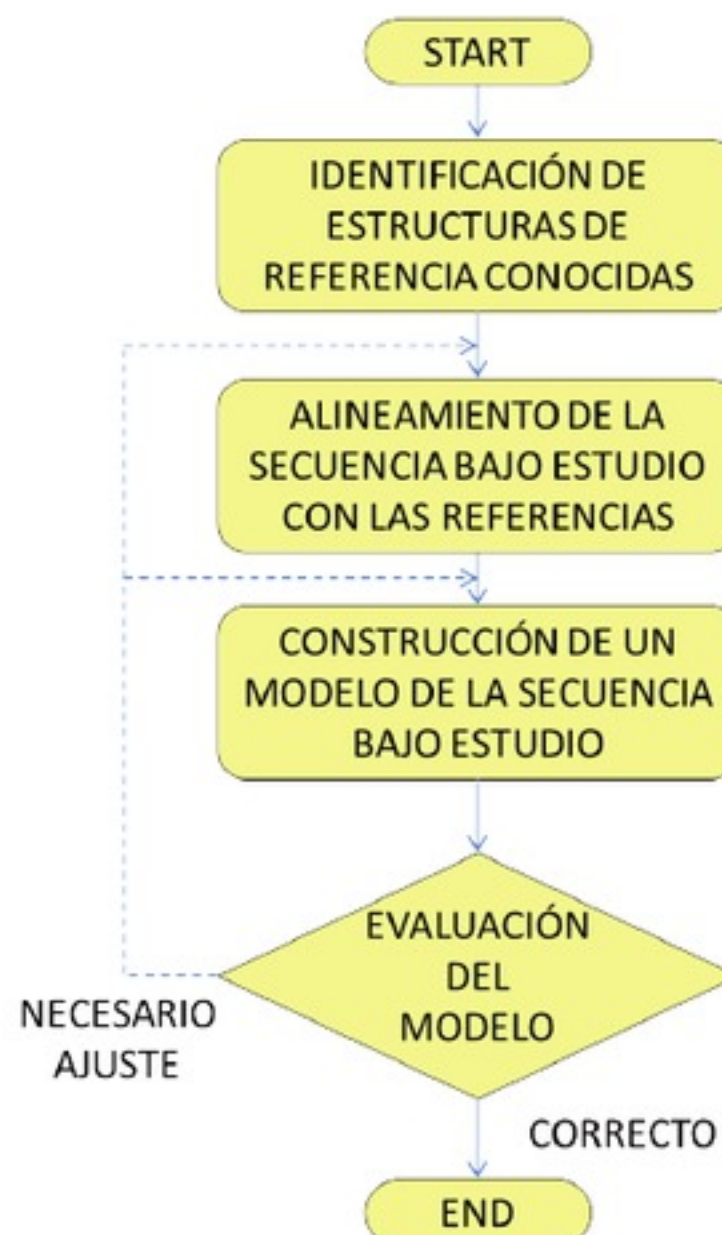


Figura 10.2. Proceso de modelado por homología

10.5 RECONOCIMIENTO DE PLIEGUES

Los métodos de reconocimiento de pliegues, en lugar de alinear secuencias, alinean estructuras. Colocan la secuencia de la proteína bajo estudio según la forma de la proteína de referencia. Por su complejidad, quedan fuera del ámbito de esta obra aunque, por completitud, hemos creído necesario reseñar su existencia.



Afortunadamente, existen herramientas informáticas que nos asisten en la predicción de la estructura 3D de proteínas utilizando métodos basados en el reconocimiento de pliegues. Una de ellas es Libellula, basada en redes neuronales y de la que puede hallarse más información en <http://www.bioinformatics.org/wiki/LIBELLULA>.



PRÁCTICA 4: ANÁLISIS DE PROTEÍNAS

A la hora de modelar una proteína bajo estudio, conviene llevar a cabo un análisis básico de su secuencia y estudiar el grado de desorden, los motivos funcionales o la ubicación subcelular, entre otros aspectos.



La calidad de la predicción aumenta si se consideran las secuencias de proteínas homólogas.

11.1 ANÁLISIS BLAST

El análisis BLAST de una proteína permite encontrar otra proteína similar en una base de datos de secuencias.

Objetivo	Herramienta BLAST
Información sobre la función de la proteína bajo estudio	BLASTP
Búsqueda de nuevos genes que codifiquen proteínas simples	TBLASTN

Tabla 11.1. Herramientas BLAST

El objetivo es encontrar proteínas similares a la siguiente:

```
>gi|108710532|gb|ABF98327.1| Oxalate oxidase 2 precursor,
putative, expressed [Oryza sativa (japonica cultivar-group)]
MEHSFKTIAAGVVIVVLLQLQAPVLIRATDADPLQDFCVADLDSKVTVNGHACKPASAAGDEFLE
SSKIATGGDVNANPNNGSNVTELDVAEWPGVNTLGVSMNRVDFAPGGTNPPHVHPRATEVGIVLRG
ELLVGIIGTLDTGNYYSKVVRAGETFVIIPRGLMHFQFNVGKTEATMVVSFNSQNPQIVFVPLTL
FGSNPPIPTPVLVKALRVDAGVVELLKSKFTGGY
```

En primer lugar, abrimos la página principal de NCBI-BLAST (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) y escogemos la herramienta BLASTP (ver Figura 11.1):

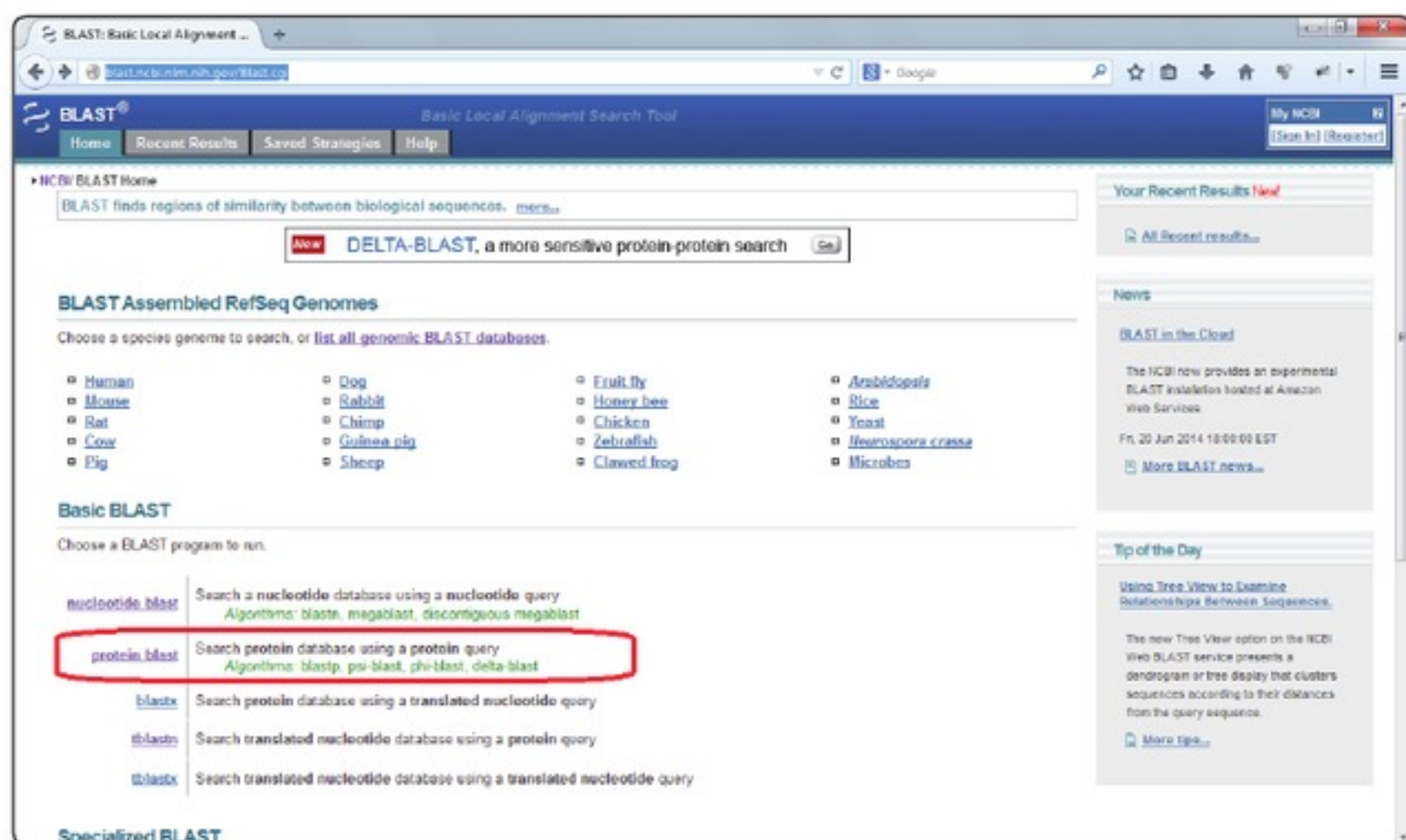


Figura 11.1. Página BLAST del NCBI

A continuación, pegamos la secuencia de la proteína en la ventana de búsqueda (ver Figura 11.2) y en **Database** escogemos la base de datos **SwissProt**.

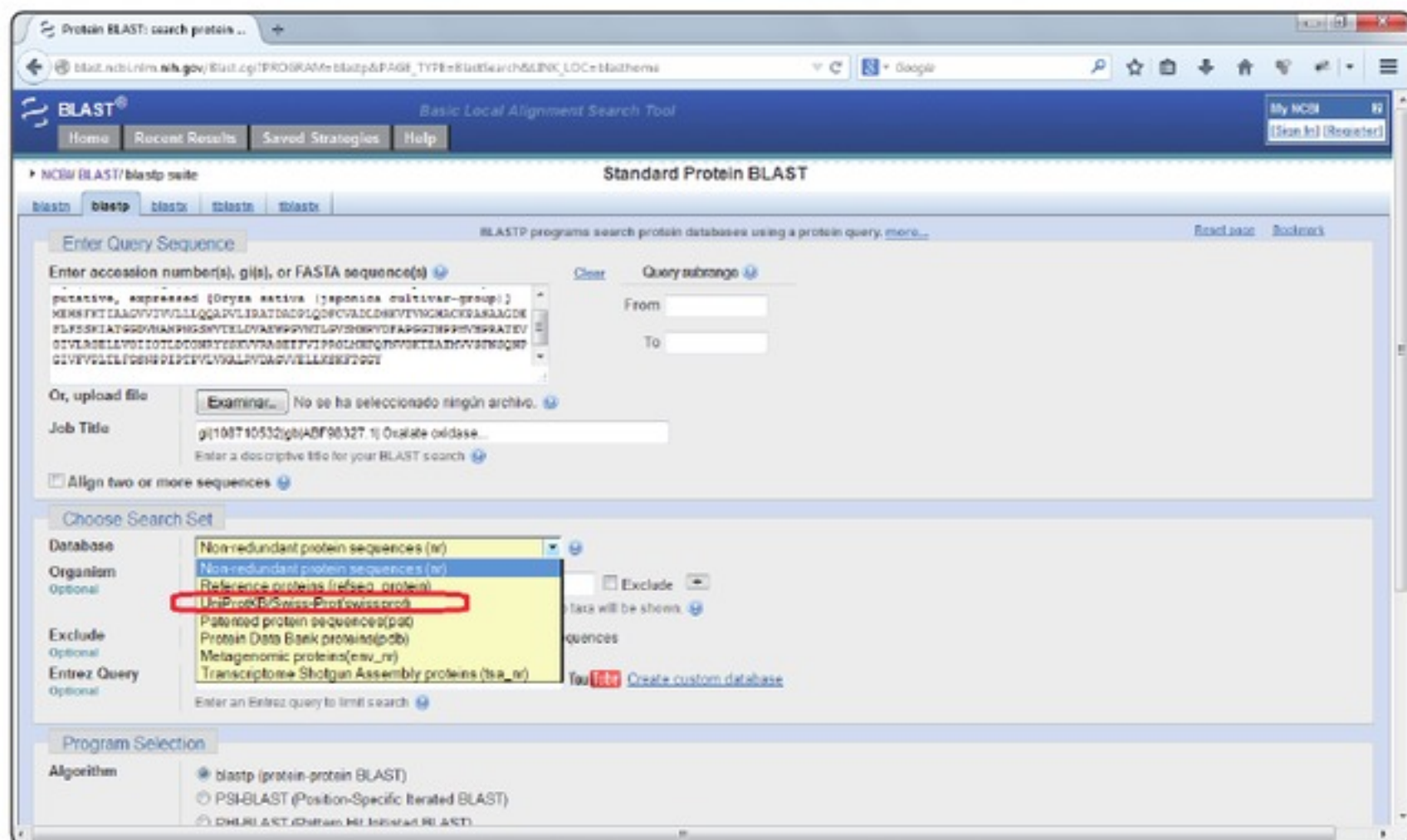


Figura 11.2. Página de análisis de la secuencia

Pinchando en el botón **BLAST**, obtenemos los resultados del análisis, que se dividen en cuatro secciones:

- Información gráfica: muestra en qué partes de la secuencia, la secuencia bajo estudio es similar a secuencias de referencia encontradas en la base de datos.
- Referencias: lista los nombres de las secuencias similares a la secuencia bajo estudio.
- Alineamientos entre la secuencia bajo estudio y todas las referencias encontradas.
- Parámetros de la búsqueda.

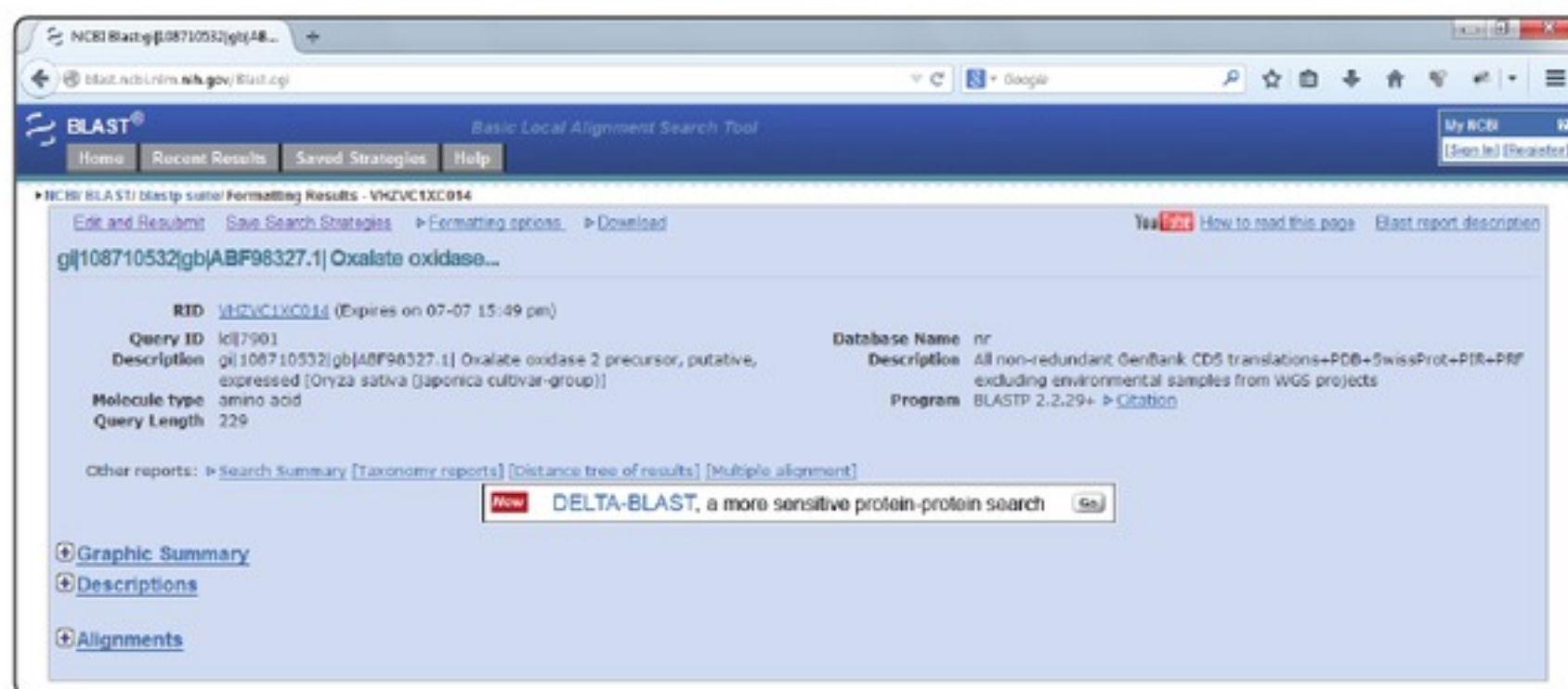


Figura 11.3. Secciones de la página de resultados de BLASTP

La sección de información gráfica ofrece un resumen visual de los resultados (ver Figura 11.4). La secuencia bajo estudio se encuentra en la parte superior y cada barra representa el fragmento de una referencia encontrada en la base de datos que coincide con la secuencia bajo estudio. Los colores indican la relevancia de la coincidencia.

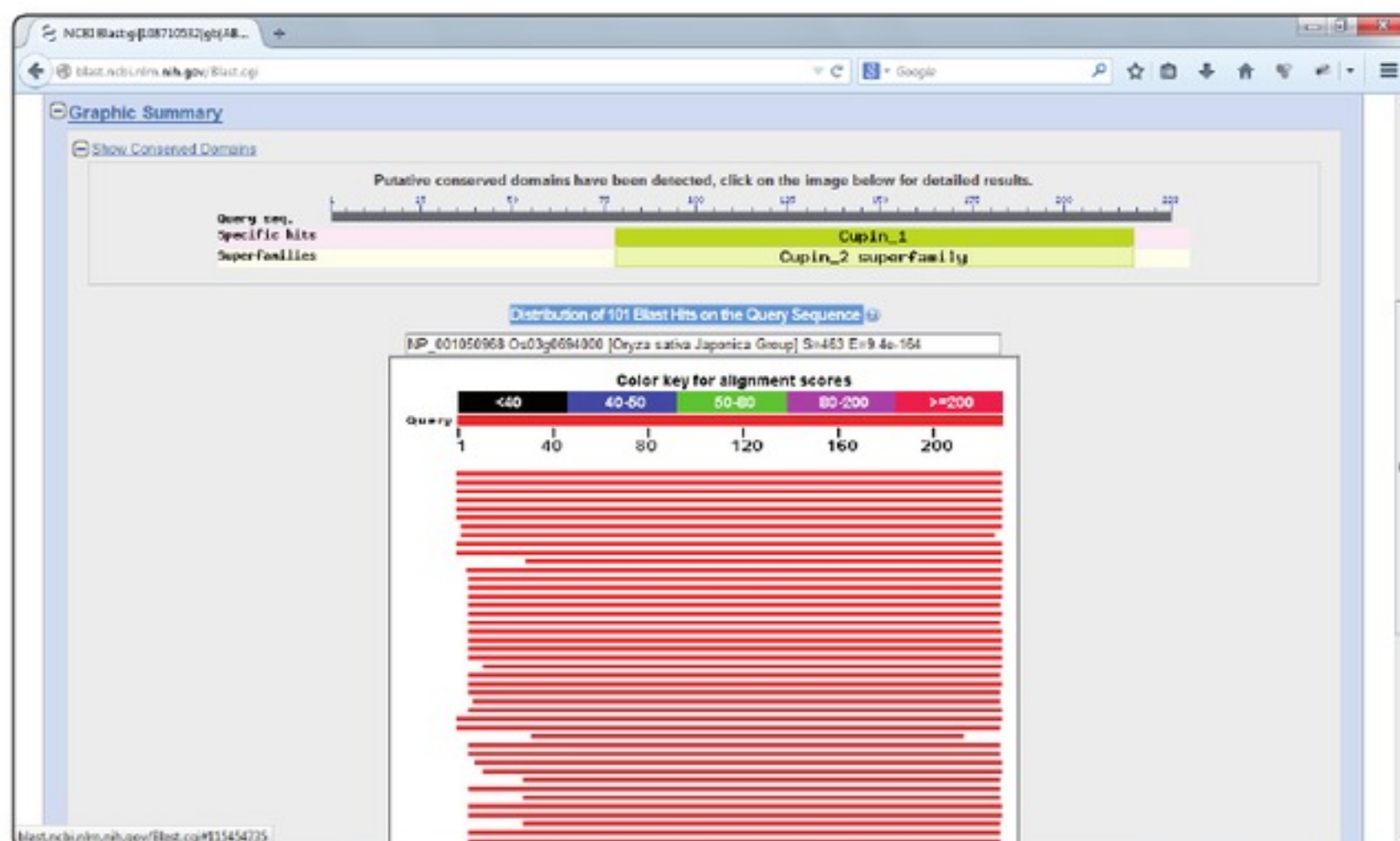
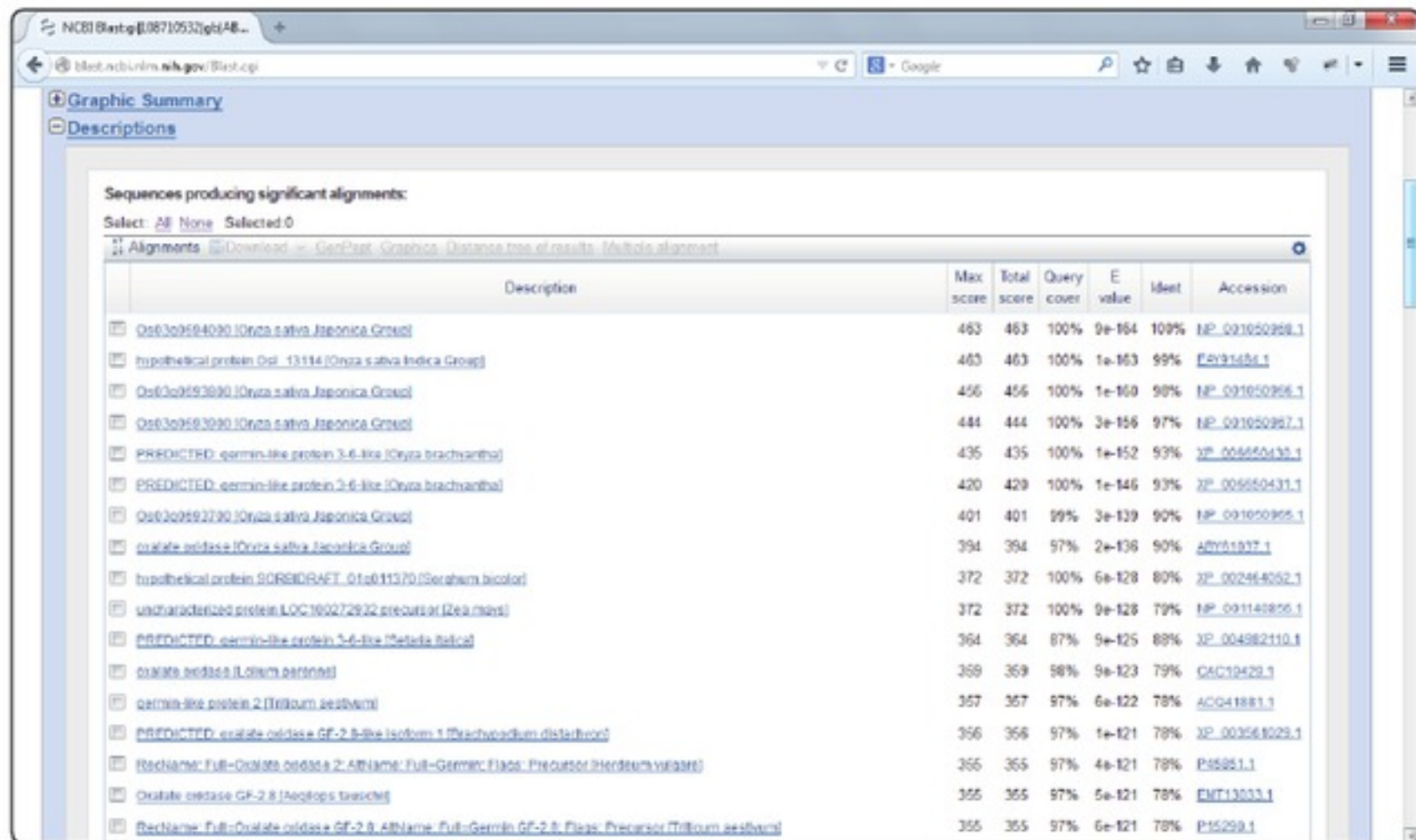


Figura 11.4. Sección de información gráfica

Por otra parte, la lista de referencias encontradas en la base de datos (ver Figura 11.5) proporciona, para cada referencia, la siguiente información:

- Nombre y *accession* de la secuencia: con un hipervínculo a la entrada de la base de datos que contiene esta secuencia.
- Descripción: procede de la anotación de la secuencia y permite saber, de un vistazo rápido, si la referencia es relevante para los fines de la investigación.
- Relevancia estadística (*bit score*) del alineamiento. Cuanto mayor sea este valor, más similares serán las secuencias.
- Valor E: es otra medida de relevancia estadística. Cuanto más similares sean las secuencias, más próximo a cero estará el valor E.



Sequences producing significant alignments:

Select: All None Selected: 0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	Ox6309584090 (Oryza sativa Japonica Group)	463	463	100%	9e-164	100%	NP_021050958.1
☐	hypothetical protein Oxi_13114 (Oryza sativa Indica Group)	463	463	100%	1e-163	99%	F59316.1
☐	Ox6309583690 (Oryza sativa Japonica Group)	456	456	100%	1e-160	98%	NP_021050956.1
☐	Ox6309583090 (Oryza sativa Japonica Group)	444	444	100%	3e-156	97%	NP_021050957.1
☐	PREDICTED: germin-like protein 3-6-like (Oryza brachyantha)	435	435	100%	1e-152	93%	XP_004455043.1
☐	PREDICTED: germin-like protein 3-6-like (Oryza brachyantha)	420	420	100%	1e-146	93%	XP_004455043.1
☐	Ox6309582790 (Oryza sativa Japonica Group)	401	401	99%	3e-139	90%	NP_021050955.1
☐	oxalate oxidase (Oryza sativa Japonica Group)	394	394	97%	2e-136	90%	AF051037.1
☐	hypothetical protein SORSDRAFT_01c011370 (Sorghum bicolor)	372	372	100%	6e-128	80%	XP_002464052.1
☐	uncharacterized protein LOC100272932 precursor (Zea mays)	372	372	100%	9e-128	79%	NP_091149856.1
☐	PREDICTED: germin-like protein 3-6-like (Setaria italica)	364	364	87%	9e-125	88%	XP_004982110.1
☐	oxalate oxidase (Lobelia Dortmunda)	369	369	98%	9e-123	79%	CAC12429.1
☐	germin-like protein 2 (Triticum aestivum)	357	357	97%	6e-122	78%	ACG41881.1
☐	PREDICTED: oxalate oxidase GF-2.8-like (Sorghum bicolor)	356	356	97%	1e-121	78%	XP_003564029.1
☐	RefName: Full-Oxalate oxidase 2; AltName: Full-Germin GF-2.8; Flag: Precursor (Sorghum bicolor)	355	355	97%	4e-121	78%	P35851.1
☐	Oxalate oxidase GF-2.8 (Aegilops triuncialis)	355	355	97%	5e-121	78%	F51303.1
☐	RefName: Full-Oxalate oxidase GF-2.8; AltName: Full-Germin GF-2.8; Flag: Precursor (Triticum aestivum)	355	355	97%	6e-121	78%	P35299.1

Figura 11.5. Referencias encontradas en la base de datos del NCBI

La Figura 11.6 muestra la información proporcionada para cada alineamiento:

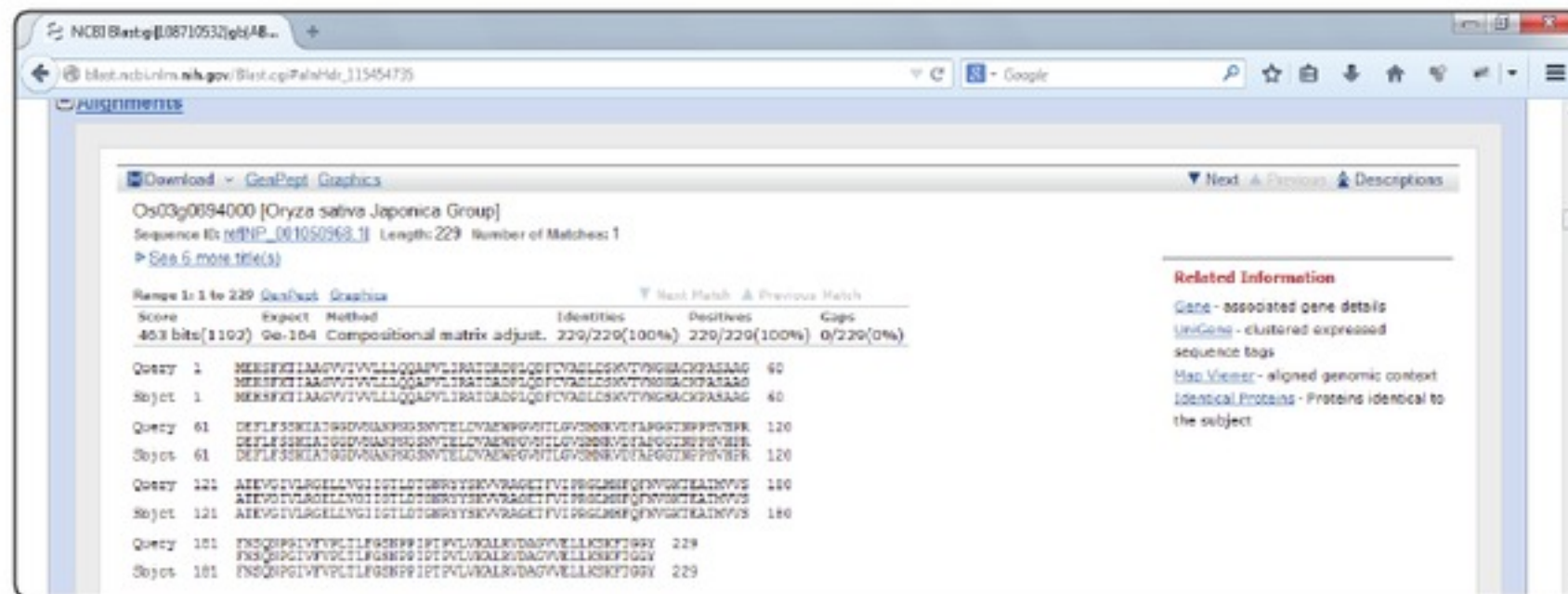


Figura 11.6. Información de un alineamiento

En primer lugar, encontramos el nombre de la referencia encontrada con la que se efectúa el alineamiento. Además, tenemos información sobre el porcentaje de similitud (un valor de 25% puede considerarse aceptable), el valor E, los positivos (porcentaje de aminoácidos que son idénticos o similares), los huecos (aminoácidos que no están alineados) y la longitud del alineamiento.

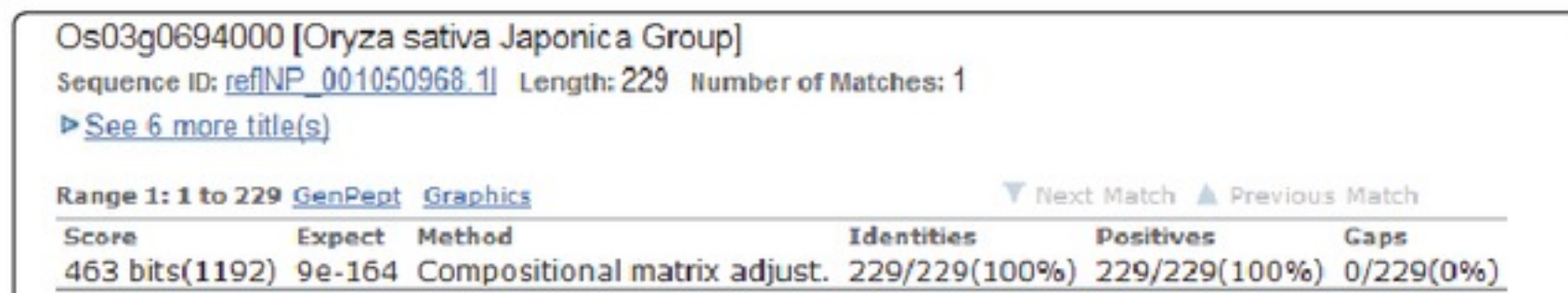


Figura 11.7. Información sobre el alineamiento

Sin embargo, la parte más interesante es que se corresponde con el alineamiento propiamente dicho (ver Figura 11.8). En la parte superior, se encuentra la secuencia bajo estudio mientras que en la parte inferior está la secuencia referencia recuperada de la base de datos. Entre ambas, aparece la línea de alineamiento, que muestra un símbolo '+' para indicar que los aminoácidos son similares; una letra, para residuos idénticos; y, un espacio para las no coincidencias. Finalmente, los números de la derecha indican las coordenadas en las que la secuencia bajo estudio y la referencia coinciden.


```

Query 1  MEHSFKTIAAGVIVVLLQLQAPVLRATDADPLQDFCVADLDSKVTVNGHACKPASAAG 60
Sbjct 1  MEHSFKTIAAGVIVVLLQLQAPVLRATDADPLQDFCVADLDSKVTVNGHACKPASAAG 60

Query 61  DEFLFSSKIATGGD/NANPNSNVTELDVAEWPGVNILGVSNNRVDFAPGGINPPHVHPR 120
Sbjct 61  DEFLFSSKIATGGD/NANPNSNVTELDVAEWPGVNILGVSNNRVDFAPGGINPPHVHPR 120

Query 121  ATEVGIVLRGELLVGIIGILDIGNRYYSK/VRAGETFVI PRGLMHFQFNVGKTEATMVVS 180
Sbjct 121  ATEVGIVLRGELLVGIIGILDIGNRYYSK/VRAGETFVI PRGLMHFQFNVGKTEATMVVS 180

Query 181  FNSQNPQIVFVPLTLFGSNPPIPTPVLVKALRVDAGVVELLKSFTGGY 229
Sbjct 181  FNSQNPQIVFVPLTLFGSNPPIPTPVLVKALRVDAGVVELLKSFTGGY 229

```

Figura 11.8. Alineamiento de secuencias

Si, por ejemplo, se hubiera escogido otra secuencia con un grado de similitud mucho más bajo, el alineamiento habría tenido el aspecto de la Figura 11.9.

Download ▾ GenPept Graphics

hypothetical protein EUGRSUZ_H04166, partial [Eucalyptus grandis]
Sequence ID: [gb|KCW61436.1](#) Length: 216 Number of Matches: 1

Range 1: 6 to 216 GenPept Graphics ▾ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
195 bits(495)	2e-58	Compositional matrix adjust.	103/213(48%)	132/213(61%)	3/213(1%)

```

Query 14  IVVLLQLQAPVLRATDADPLQDFCVADLDSKVTVNGHACK-PASAAGDEFLFSSKIATG 72
Sbjct 6   ISLLILALATATTFAVDPSPLQDFCVAINDPKVFVNGKFKDPKQVTADDFLFGFRYPG 65

Query 73  GDVNIANPNSNVTELDVAEWPGVNILGVSNNRVDFAPGGINPPHVHPRATEVGIVLRGEL 132
Sbjct 66  N--7ANPLGSKVTPAFVDQFAGNLTGISMARIDFAPGGINPPHHPRGTEILVVTEGTL 123

Query 133  LVGIIGILDIGNRYYSK/VRAGETFVI PRGLMHFQFNVGKTEATMVVSFNSQNPQIVFVP 192
Sbjct 124  LVGFVTSNQLNNTFFTKVLYKGDFVFFIGLIHFQNLNIGHTPALAFAALSSQNPGLITIA 183

Query 193  LTLFGSNPPIPTPVLVKALRVDAGVVELLKSFTGGY 225
Sbjct 184  NAVFGSKPPISAGVLTAKAFQVDEKVVVDYLQAQF 216

```

Figura 11.9. Alineamiento de secuencias

11.2 BÚSQUEDA DE DOMINIOS FUNCIONALES

Una proteína es una unidad muy compleja con gran cantidad de funciones. Para simplificar el análisis de estas funciones, se definen los llamados dominios funcionales encargados, cada uno de ellos, de una función molecular concreta.



En esta parte de la práctica estudiaremos cómo buscar dominios funcionales en las bases de datos más empleadas (EBI-Interpro y PFAM), pero hay más. Sugerimos al lector que repita los ejemplos con HHPred (<http://hhpred.tuebingen.mpg.de/hhpred>) o DomPred (<http://bioinf.cs.ucl.ac.uk/dompred>) y compare los resultados.

11.2.1 Búsqueda de dominios con EBI-Interpro

Supongamos que debemos encontrar el dominio funcional y la familia de una secuencia proteica desconocida.

```
>Proteina bajo estudio
MEYGFKAAGLVFVLLQLQAPVLIRATDADPLQDFCVADLNSEVTVNGHACKPASAAAGDEFLEFS
KIATGGDVNANPNNGSNVTELDVAEWPGVNTLGVSMMNRVDFAPGGTNPPHVHPRATEVGIVLRGEL
LVGIIGTLDTGNRYYSKVVRAGETFVIPRGLMHFQFNVGKTEATMVVSFNSQNPQIVFVPLTLFG
SNPPIPTPVLVKALRVDAGVVELLKSKFTGGY
```



Un dominio es un fragmento de la secuencia que tiene una estructura estable y muestra un nivel de conservación a lo largo de la evolución de la proteína.

Este problema puede resolverse auxiliándonos con la herramienta EBI-InterPro (<http://www.ebi.ac.uk/interpro/>).

Una vez abierta la página principal de InterPro, se copia la secuencia de proteínas bajo estudio en el área de texto acondicionada para ello (ver Figura 11.10):

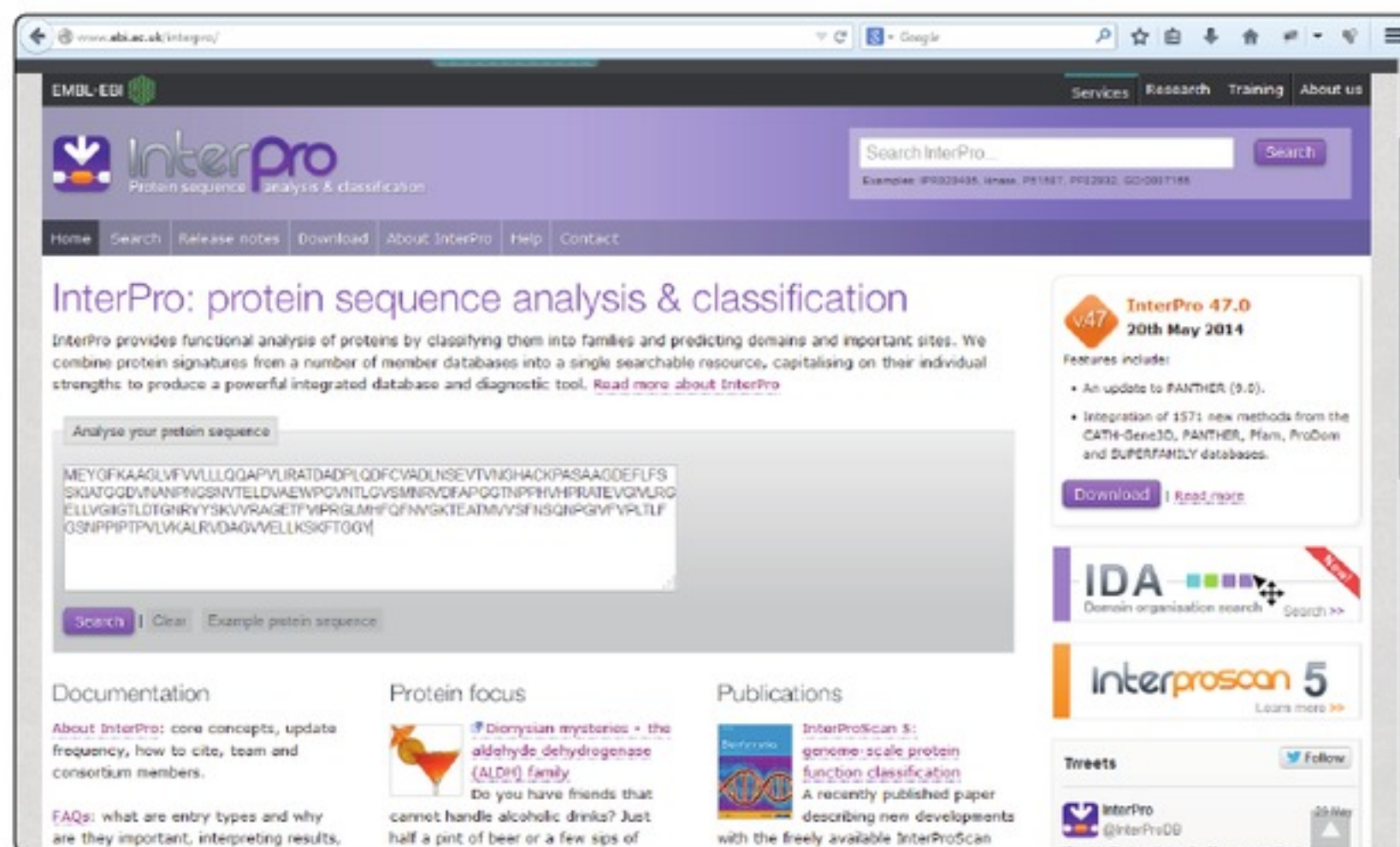


Figura 11.10. Página principal de InterPro

Tras pinchar en el botón **Search**, InterPro muestra la pantalla de procesamiento del trabajo (ver Figura 11.11).

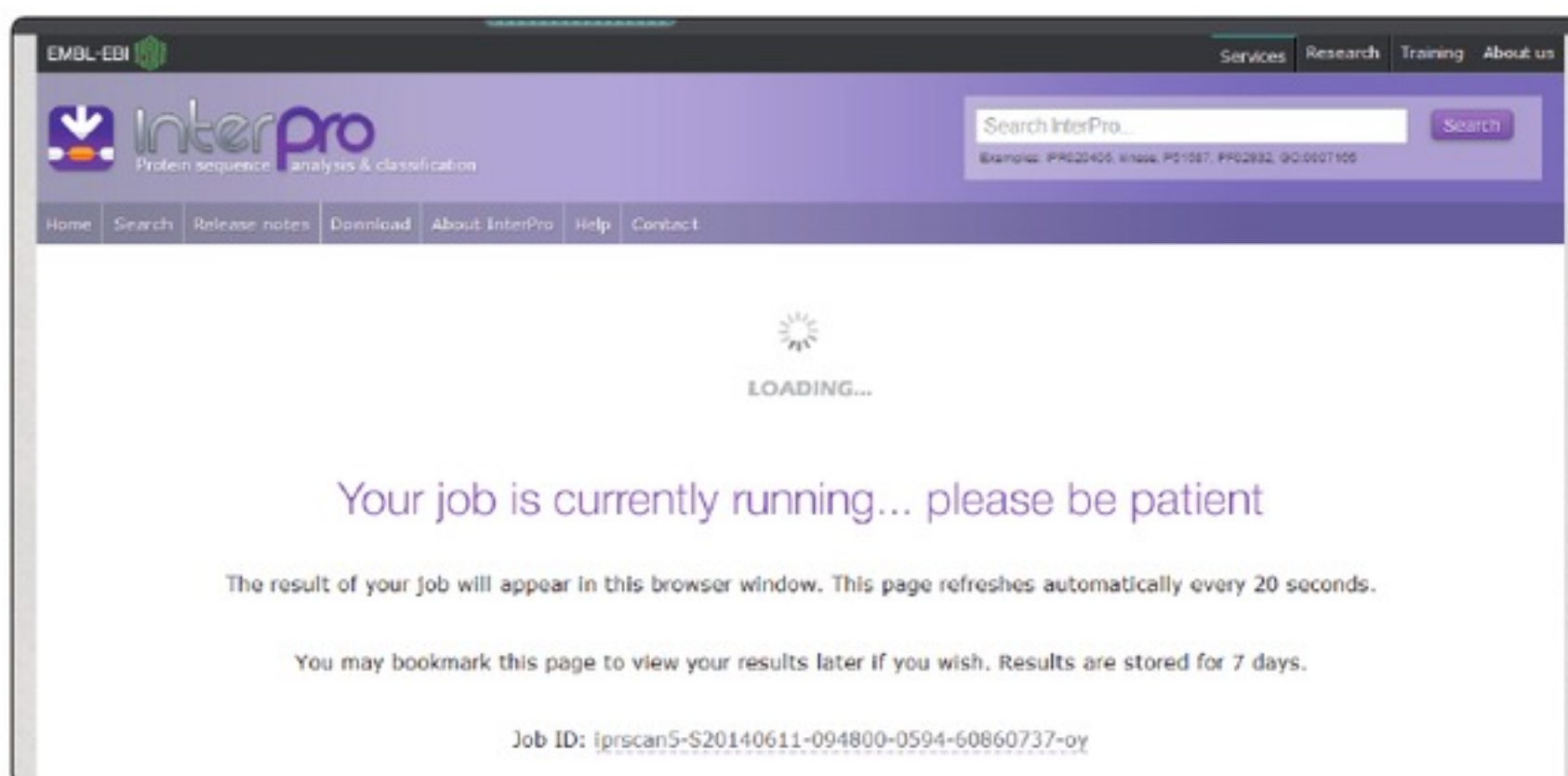


Figura 11.11. Procesamiento del trabajo

Una vez que el procesamiento del trabajo ha finalizado, aparece la página de resultados. Esta página dispone de una sección de filtrado en donde es posible especificar el tipo de resultados que se desean ver según una serie de criterios. El más importante de ellos es el tipo de entrada. A cada entrada de InterPro se le asigna uno de los siguientes tipos, que permiten inferir cuándo una proteína coincide con una entrada:

- **Familia (*Family*):** una familia de proteínas en un grupo de proteínas que comparten una evolución común, lo que queda reflejado en que desempeñan funciones relacionadas o tienen una estructura primaria, secundaria o terciaria similar.
- **Dominio (*Domain*):** los dominios pueden existir en una amplia variedad de contextos biológicos y se caracterizan por una estructura, función o fragmento de secuencia.
- **Repetición (*Repeat*):** una coincidencia con una entrada de este tipo indica que la secuencia bajo estudio presenta fragmentos que se repiten en la proteína con la que coincide.
- **Sitio (*Site*):** son secuencias cortas que contienen una o más regiones conservadas.

Teniendo en cuenta lo anterior, el resultado obtenido es que la secuencia pertenece a la familia GERMIN (IPR001929), que se han encontrado coincidencias en los dominios IPR014710, IPR011051 e IPR006045 así como un *site* IPR019780.

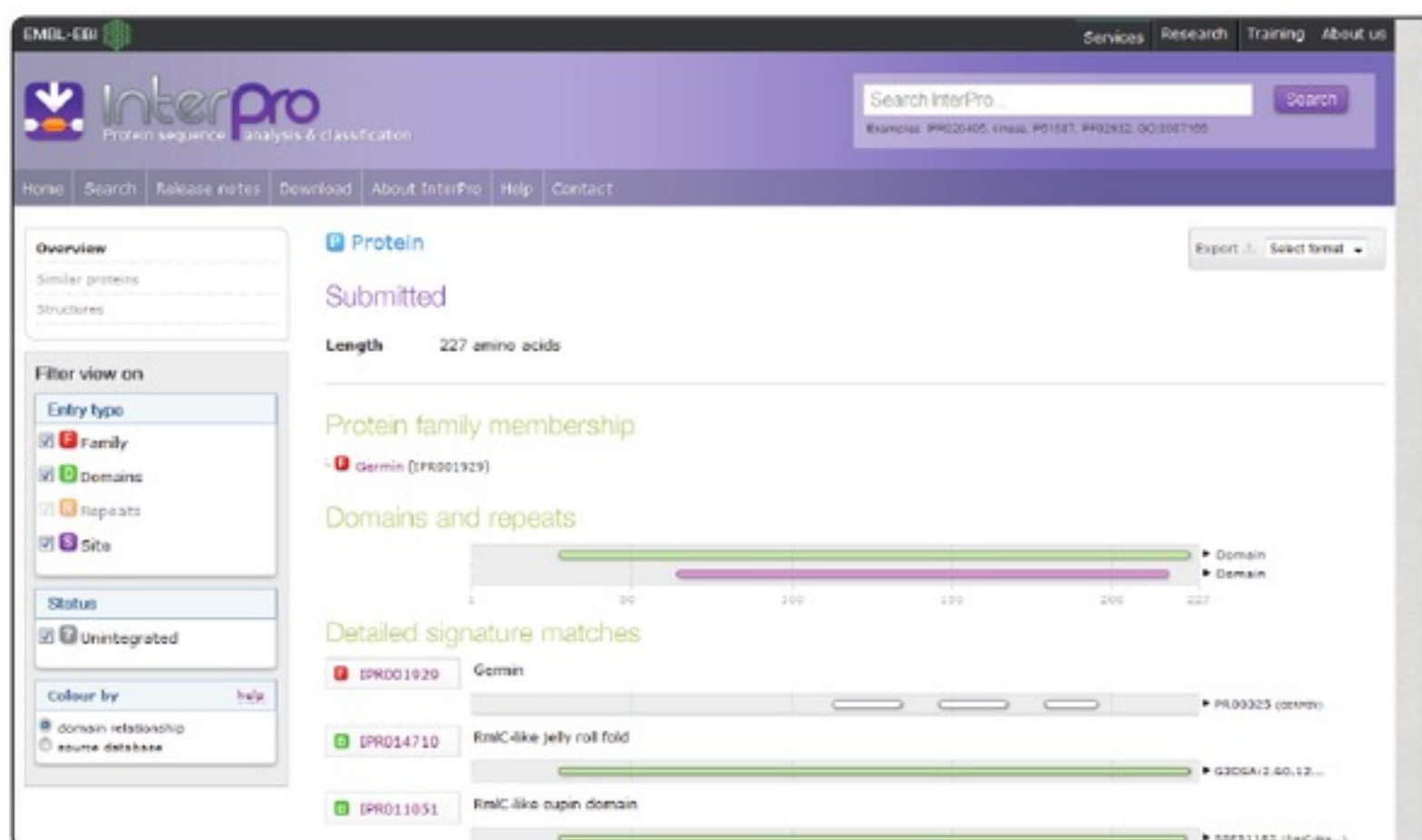


Figura 11.12. Resultado del análisis de la secuencia

Además, también nos predice las funciones que tendrá la proteína correspondiente a la secuencia bajo estudio:



Figura 11.13. Predicción de las funciones de la proteína

Por otro lado, es posible exportar los resultados en varios formatos y consultar las bases de datos biológicas de las que se ha obtenido la información.

Si pinchamos, por ejemplo, en el nombre de la familia, en otra ventana aparecerá información relativa a la misma (ver Figura 11.14):

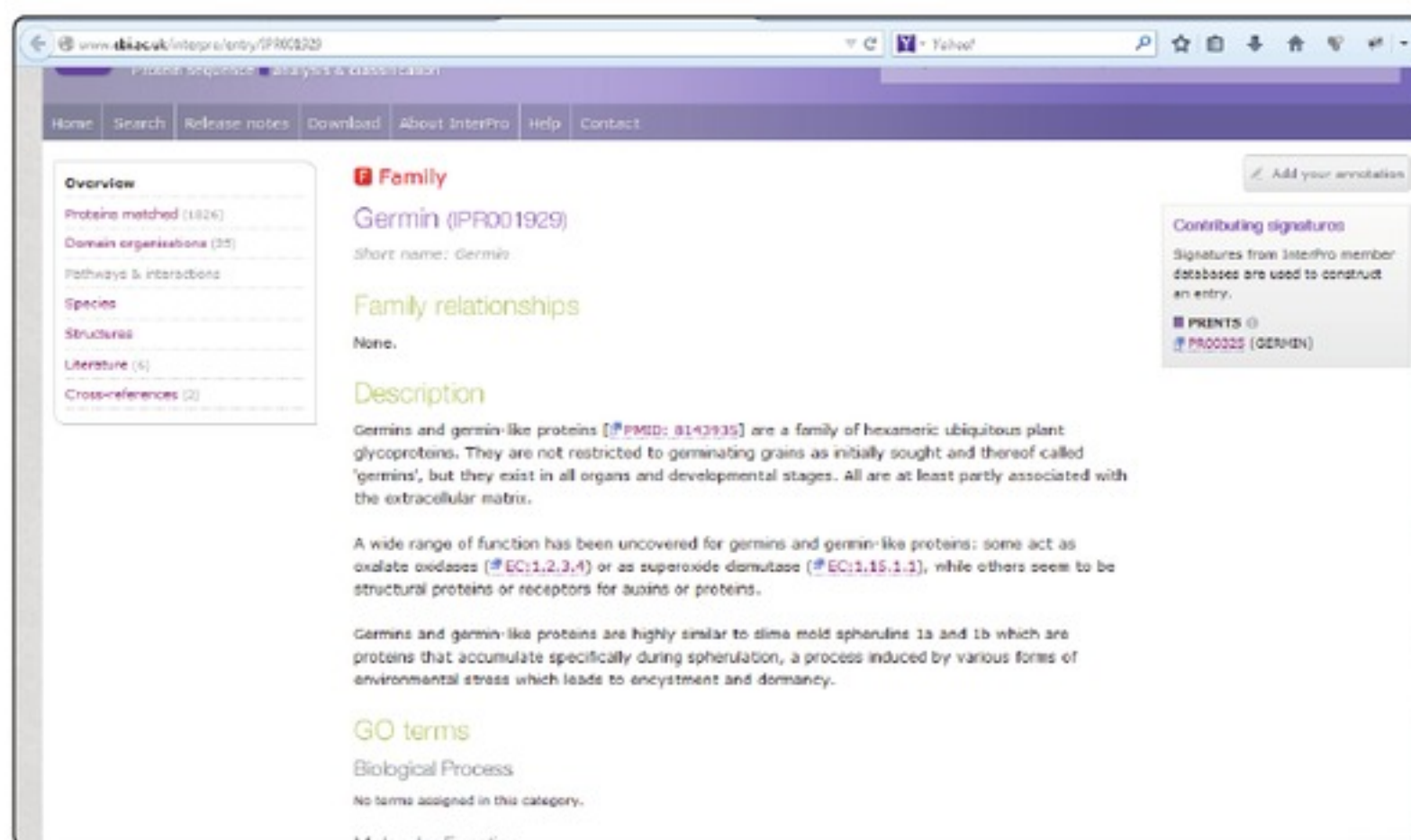


Figura 11.14. Información sobre la familia de proteínas

11.2.2 Búsqueda de dominios con PFAM

PFAM (<http://pfam.xfam.org/>) se basa en alineamientos múltiples para definir familias de secuencias o dominios y facilita su localización en otras proteínas. Dentro de PFAM existen dos niveles de calidad:

- PFAM-A, que son entradas curadas de alta calidad y que cubren un amplio rango de la base de datos de secuencias.
- PFAM-B, se trata de entradas generadas automáticamente a partir del procesamiento de la información contenida en la base de datos ADDA (*Automatic Domain Decomposition Algorithm*), accesible en http://ekhidna.biocenter.helsinki.fi/sqgraph/pairsdb/index_html. Aunque de menor calidad que la anterior, las familias PFAM-B son útiles para identificar regiones conservadas funcionalmente que no se encuentran en PFAM-A.

Una entrada PFAM se clasifica en una de cuatro maneras: familia (colección de regiones de proteínas relacionadas), dominio (unidad estructural), repetición (unidad corta que es inestable de manera aislada pero forma una estructura estable cuando está presente más de una copia) y motivo (unidad corta que se puede encontrar fuera de dominios globulares). Por otra parte, algunas familias PFAM se agrupan en clanes o colecciones de familias que han evolucionado de un origen común y que se caracterizan por presentar una estructura terciaria similar o, cuando las estructuras no están disponibles, motivos comunes en sus respectivas secuencias.

Al igual que en el caso anterior, trabajaremos sobre la siguiente secuencia:

```
>Proteina bajo estudio
MEYGFKAAGLVFVLLQLQAPVLIRATDADPLQDFCVADLNSEVTVNGHACKPASAAGDEFLFSS
KIATGGDVNANPNNGSNVTELDVAEWPGVNTLGVSMMNRVDFAPGGTNPPHVHPRATEVGIVLRGEL
LVGIIIGTLDGTGNRYYSKVVRAGETFVIPRGLMHFQFNVGKTEATMVVSFNSQNPGLVFPVPLTLFG
SNPPIPTPVLVKALRVDAGVVELLKSKFTGGY
```

Figura 11.15. Secuencia de la proteína bajo estudio

En la página principal de PFAM (ver Figura 11.16), seleccionamos la opción **Sequence Search**. Aparecerá una caja de texto en donde pegaremos la secuencia bajo estudio (ver Figura 11.17) y pincharemos en el botón **Go** para proceder a la búsqueda.

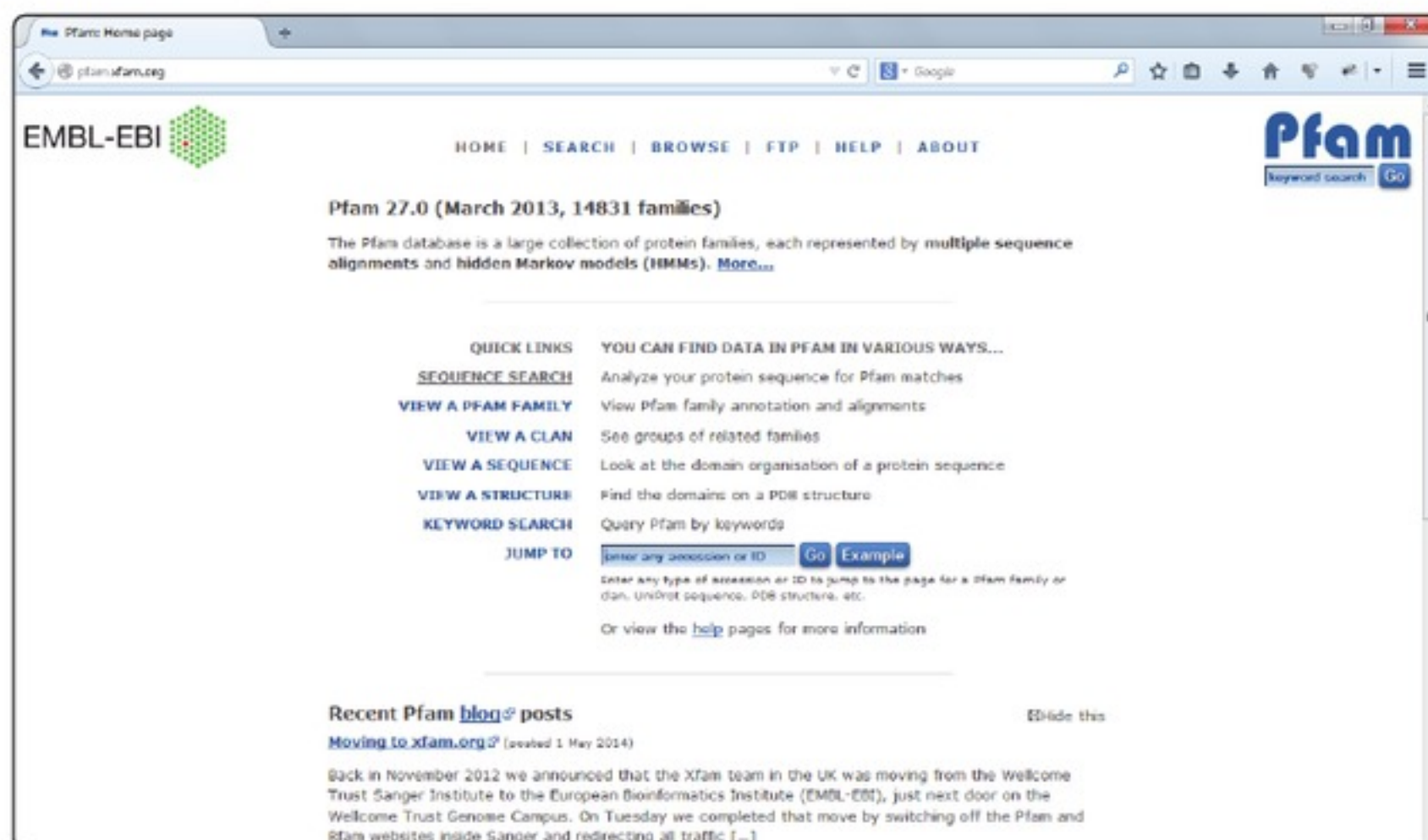


Figura 11.16. Página principal de PFAM

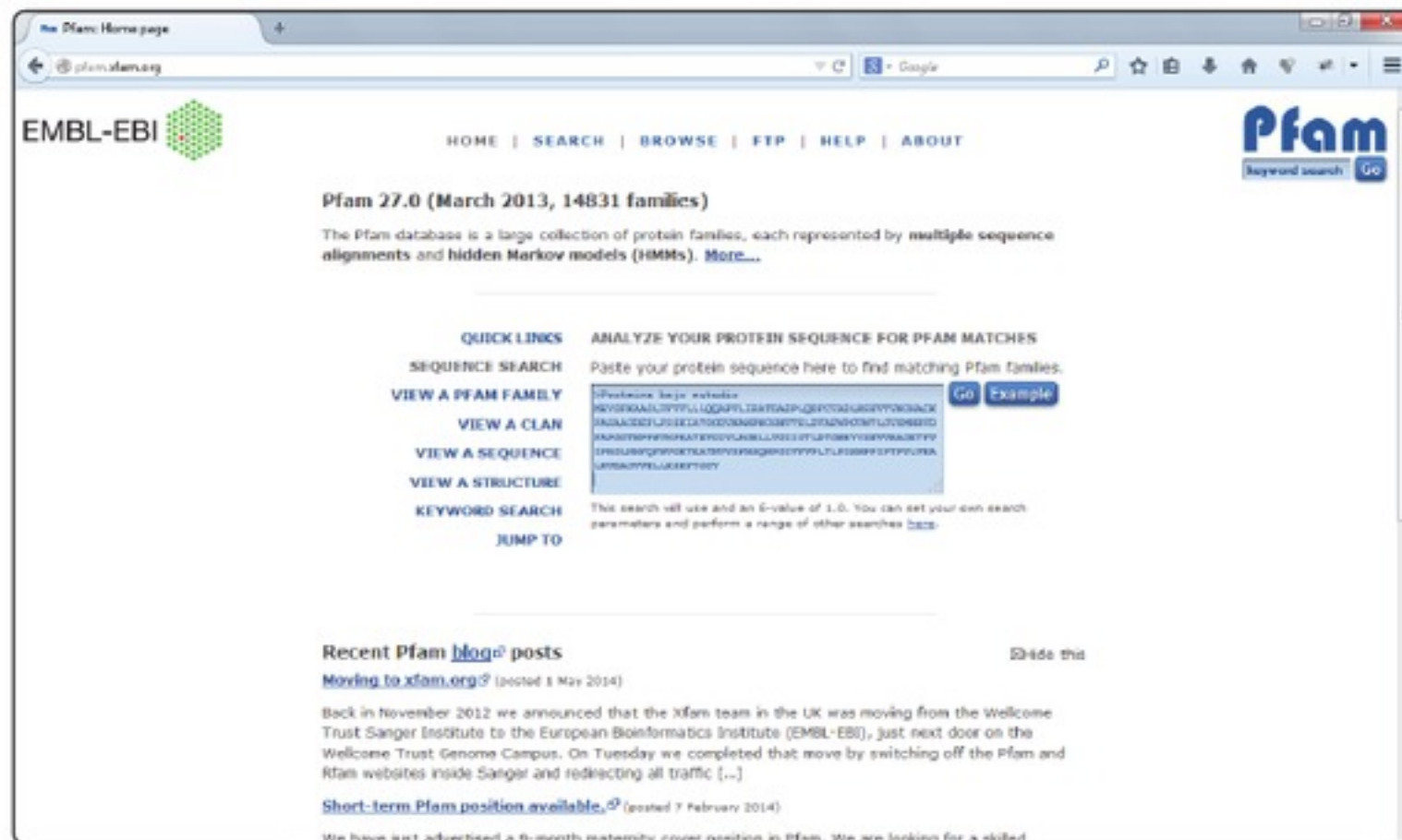


Figura 11.17. Análisis de secuencias con PFAM

Si queremos incluir las familias PFAM-B en la búsqueda, hay que pinchar en el enlace **Keywords search** y se abrirá la siguiente pantalla (ver Figura 11.18):

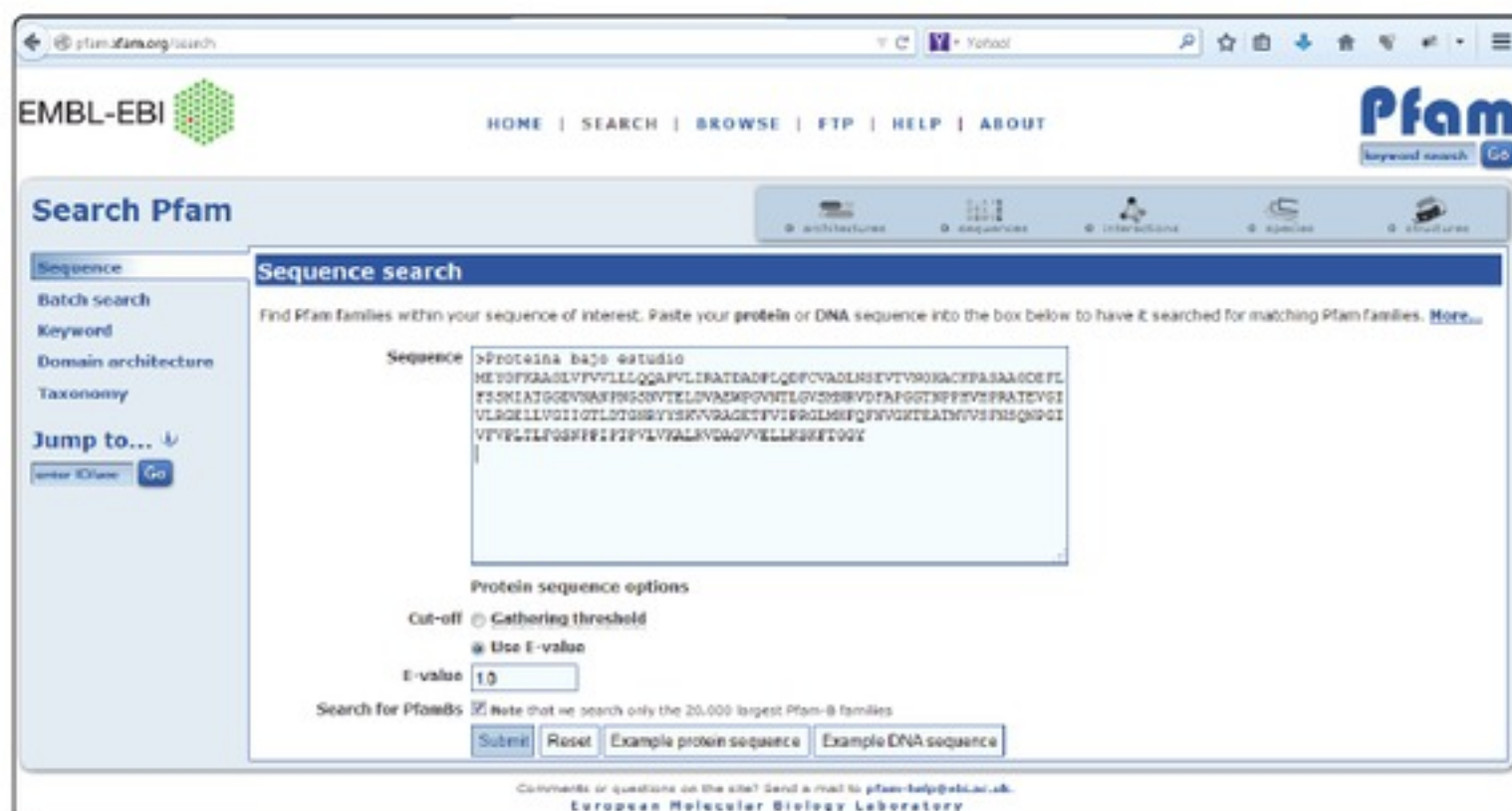


Figura 11.18. Opciones de búsqueda

Aparecerá, entonces, la página de resultados (ver Figura 11.19), en donde encontramos, coincidiendo con los obtenidos en el epígrafe anterior, que la proteína bajo estudio pertenece al dominio funcional Cupin_1.

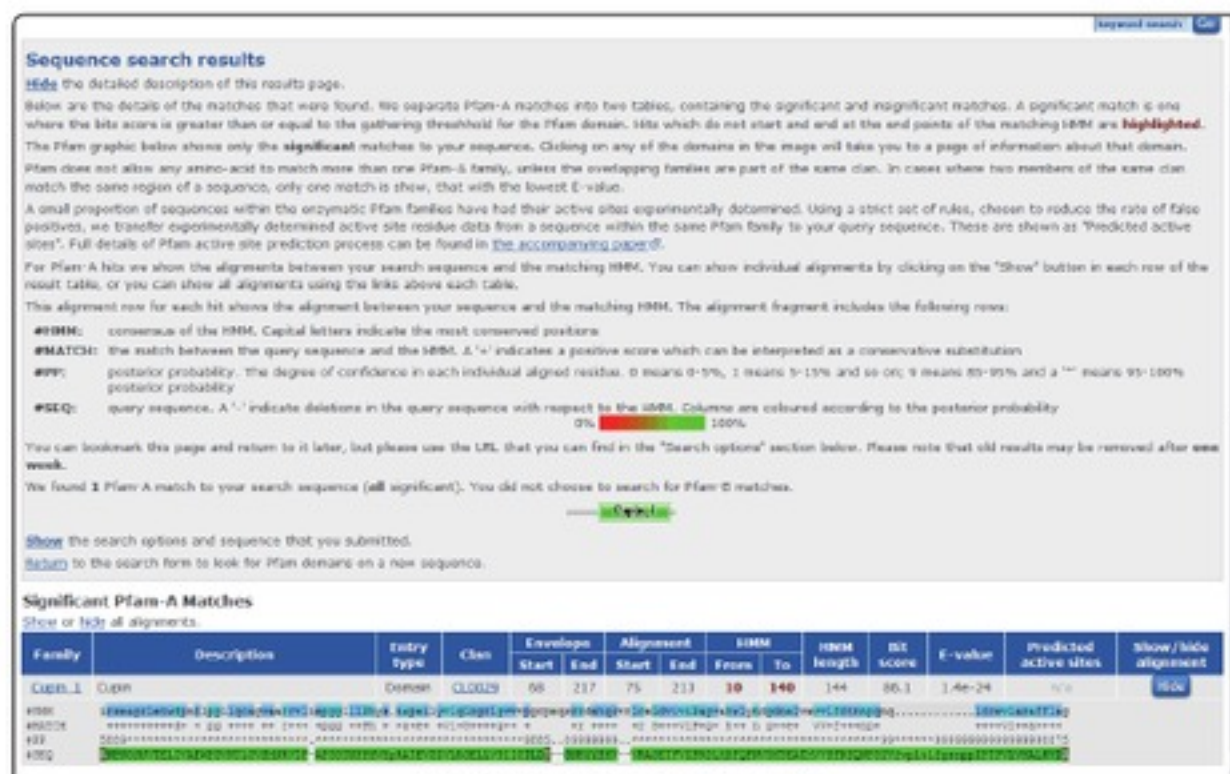


Figura 11.19. Resultados del análisis

Una de las posibilidades que ofrece PFAM es consultar directamente las características de los dominios funcionales obtenidos en el análisis. Si pinchamos en el enlace de la columna **Family**, PFAM mostrará en el navegador las anotaciones disponibles sobre el dominio funcional en cuestión (ver Figura 11.20).

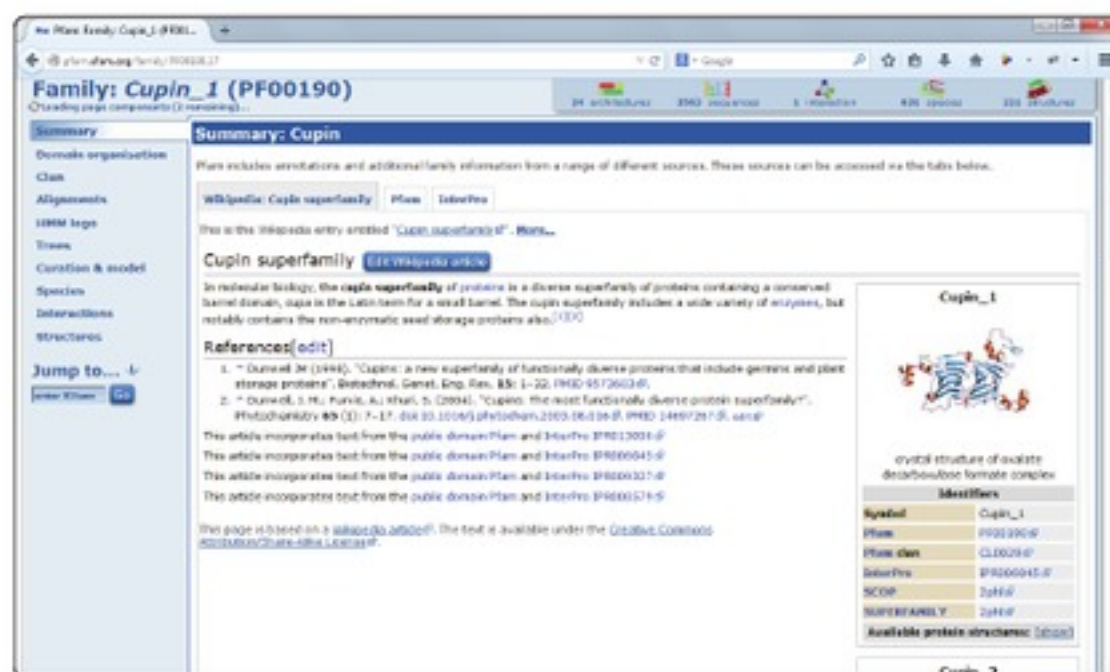


Figura 11.20. Anotaciones sobre el dominio funcional de la proteína bajo estudio

11.3 PREDICCIÓN DE LA UBICACIÓN SUBCELULAR

El análisis subcelular del proteoma es un proceso que incluye, entre otros aspectos, la predicción del destino de la proteína bajo estudio en la célula a partir de la secuencia de aminoácidos de la misma, lo que constituye un paso importante para determinar su función bioquímica.

Según el dogma central de la Biología Molecular, la copia original del mensaje que porta la proteína se encuentra en el ADN y se transcribe en el ARNm para ser transportado y, posteriormente, traducido con el fin de sintetizar la proteína. A partir de ese momento, ocurren ciertos procesos postraduccionales previos a la utilización de la proteína por parte de la célula, ya que debe ser transportada desde los ribosomas (donde se sintetiza) hasta el destino final.

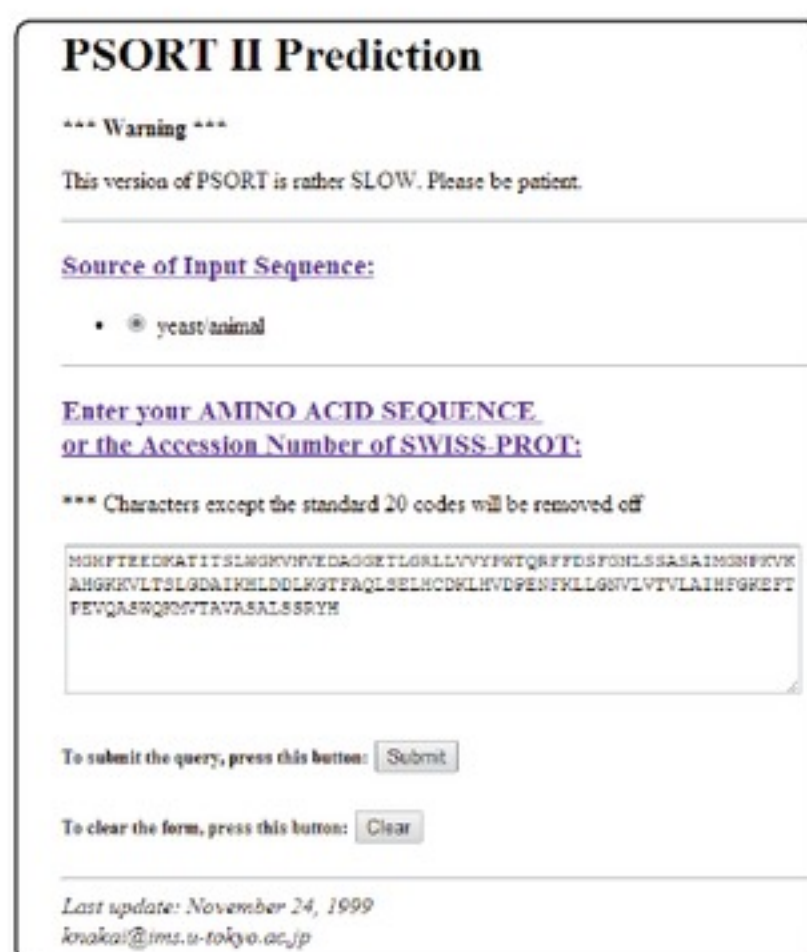
Afortunadamente, es posible predecir la ubicación final de una proteína con la utilización de herramientas informáticas. Una de estas herramientas es PSORT (<http://psort.hgc.jp/>).

En esta práctica predeciremos la ubicación subcelular de la hemoglobina humana subunidad gamma-1, cuyo *accession* en UniProt es P69891 y cuya secuencia de aminoácidos es la de la Figura 11.21.

```
>gi|56749860|sp|P69891.2|HBG1_HUMAN RecName: Full=Hemoglobin
subunit gamma-1; AltName: Full=Gamma-1-globin; AltName: Full=Hb
F Agamma; AltName: Full=Hemoglobin gamma-1 chain; AltName:
Full=Hemoglobin gamma-A chain
MGHFTTEEDKATITSLWGKVNVEDAGGETLGRLLVVPWTQRFFDSFGNLSSASAIMGNPKVKAHG
KKVLTSLGDAIKHLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLTVLAIHFGKEFTPEVQAS
WQKMVTAVASALSSRYH
```

Figura 11.21. Proteína bajo estudio

En la página principal de PSORT, introducimos la secuencia anterior y pulsamos el botón **Submit**:



PSORT II Prediction

*** Warning ***
This version of PSORT is rather SLOW. Please be patient.

Source of Input Sequence:

- ☒ yeast/animal

Enter your AMINO ACID SEQUENCE
or the Accession Number of SWISS-PROT:

*** Characters except the standard 20 codes will be removed off

MGHFTTEEDKATITSLWGKVNVEDAGGETLGRLLVVPWTQRFFDSFGNLSSASAIMGNPKVKAHG
KKVLTSLGDAIKHLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLTVLAIHFGKEFTPEVQAS
WQKMVTAVASALSSRYH

To submit the query, press this button:

To clear the form, press this button:

Last update: November 24, 1999
knakai@ims.u-tokyo.ac.jp

Figura 11.22. Introducción de datos de la búsqueda

Una vez que el trabajo se ha procesado, aparece la página de resultados (ver Figura 11.23).



Figura 11.23. Página de resultados de PSORT

PSORT ejecuta, para la misma secuencia proteica, una variedad de programas orientados cada uno de ellos a determinar diferentes propiedades de dicha secuencia. Entrar en una descripción detallada de dichas propiedades queda fuera del ámbito de esta publicación, aunque puede encontrarse en el manual de ayuda de la herramienta.

11.4 BÚSQUEDA DE ESTRUCTURAS DE REFERENCIA

En el capítulo 10 introdujimos los principios fundamentales del modelado de proteínas. Para poner en práctica los conceptos explicados allí, trabajaremos con la siguiente secuencia:

```
>gi|108710532|gb|ABF98327.1| Oxalate oxidase 2 precursor,
putative, expressed [Oryza sativa (japonica cultivar-group)]
MEHSFKTIAAGVVIVVLLQLQAPVLIRATDADPLQDFCVADLDSKVTVNGHACKPASAAGDEFLE
SSKIATGGDVNANPNNGSNVTELDVAEWPVNTLGVSMNRVDFAPGGTNPPHVHPRATEVGIVLRG
ELLVGIIGTLDTGNNRYYSKVVRAGETFVIPRGLMHFQFNVGKTEATMVVSFNSQNPNGIVFVPLTL
FGSNPPIPTPVLVKALRVDAGVVVELLKSFTGGY
```

Figura 11.24. Proteína bajo estudio

A la hora de modelar y predecir la estructura y funciones de una proteína, la mejor fuente de información es el PDB. En la página principal de PDB, seleccionamos la opción de búsqueda avanzada. Aparecerá un desplegable con una amplia variedad de opciones que determinan el método de búsqueda. Puesto que disponemos de la secuencia en formato FASTA, seleccionaremos la opción **Sequence (BLAST/FASTA/PSI-BLAST)** y pegamos la secuencia bajo estudio en el área de texto disponible (ver Figura 11.25). Además de la secuencia podemos especificar la herramienta de búsqueda (BLAST, FASTA o PSI-BLAST), el valor E de umbral por encima del cual no se considerará representativa la coincidencia encontrada en la base de datos o el porcentaje de similitud que marca el límite de significado de la coincidencia en la base de datos.

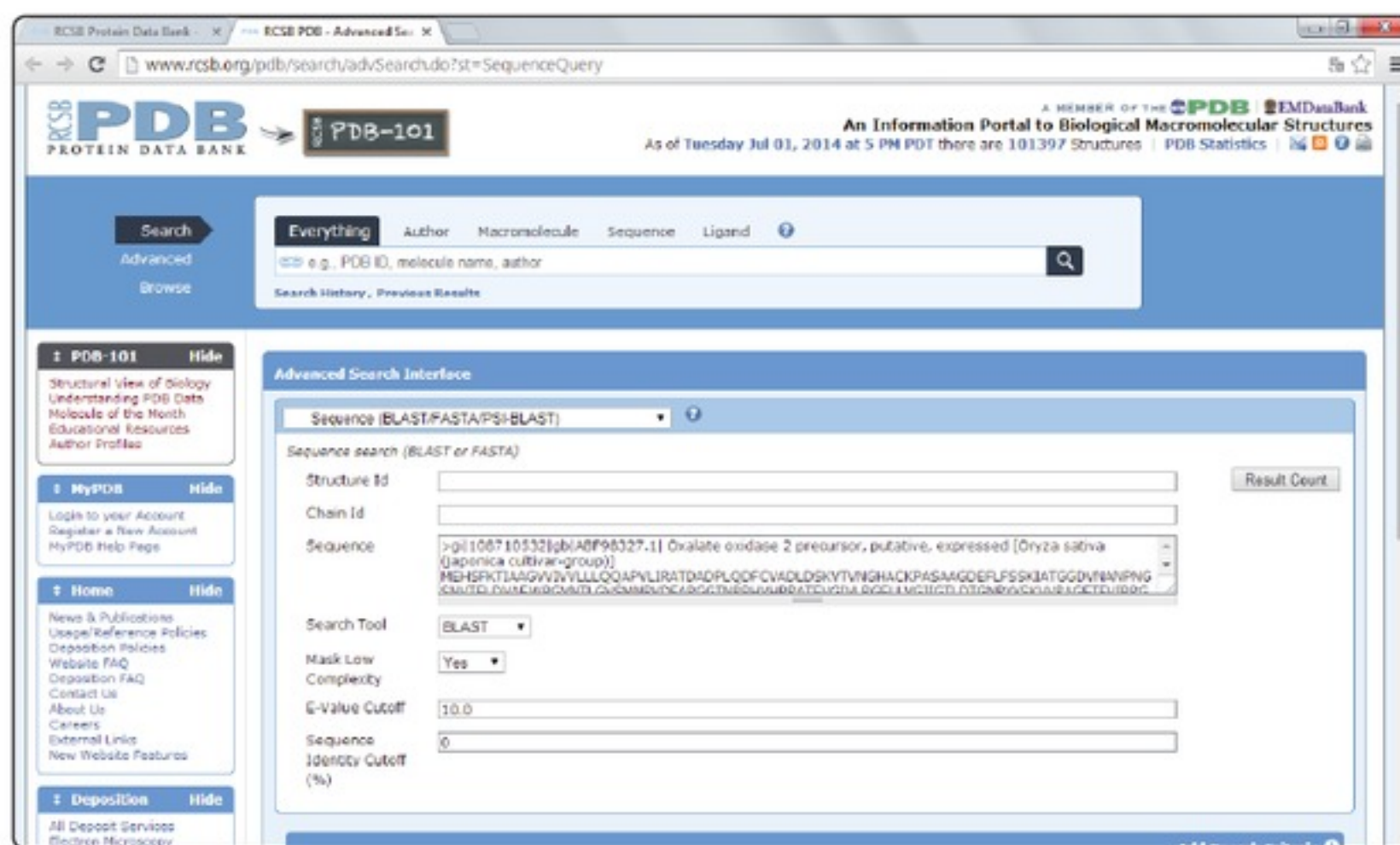


Figura 11.25. Búsqueda por secuencia en PDB

En cualquier momento es posible conocer el número de resultados de la búsqueda pulsando en el botón **Result Count** (ver Figura 11.26):



Figura 11.26. Recuento de resultados

Es posible añadir más criterios de búsqueda y relacionarlos entre sí pulsando en el enlace **Add Search Criteria** de la parte derecha de la pantalla (ver Figura 11.27):



Figura 11.27. Adición de criterios de búsqueda

Finalmente, si pulsamos en el botón **Submit Query**, obtenemos la página de resultados (ver Figura 11.28) en la que encontramos tres secciones: entradas en la base de datos de estructuras, citas académicas y científicas y entradas en la base de datos de ligandos. Explicaremos cada una de estas secciones por separado.

En la primera pestaña encontramos los resultados de la consulta realizada en la base de datos de estructuras, con una sección en la que es posible refinar los resultados de dicha búsqueda, incluso de manera gráfica. Debajo de estas opciones de búsqueda avanzada, se encuentra la lista de coincidencias.

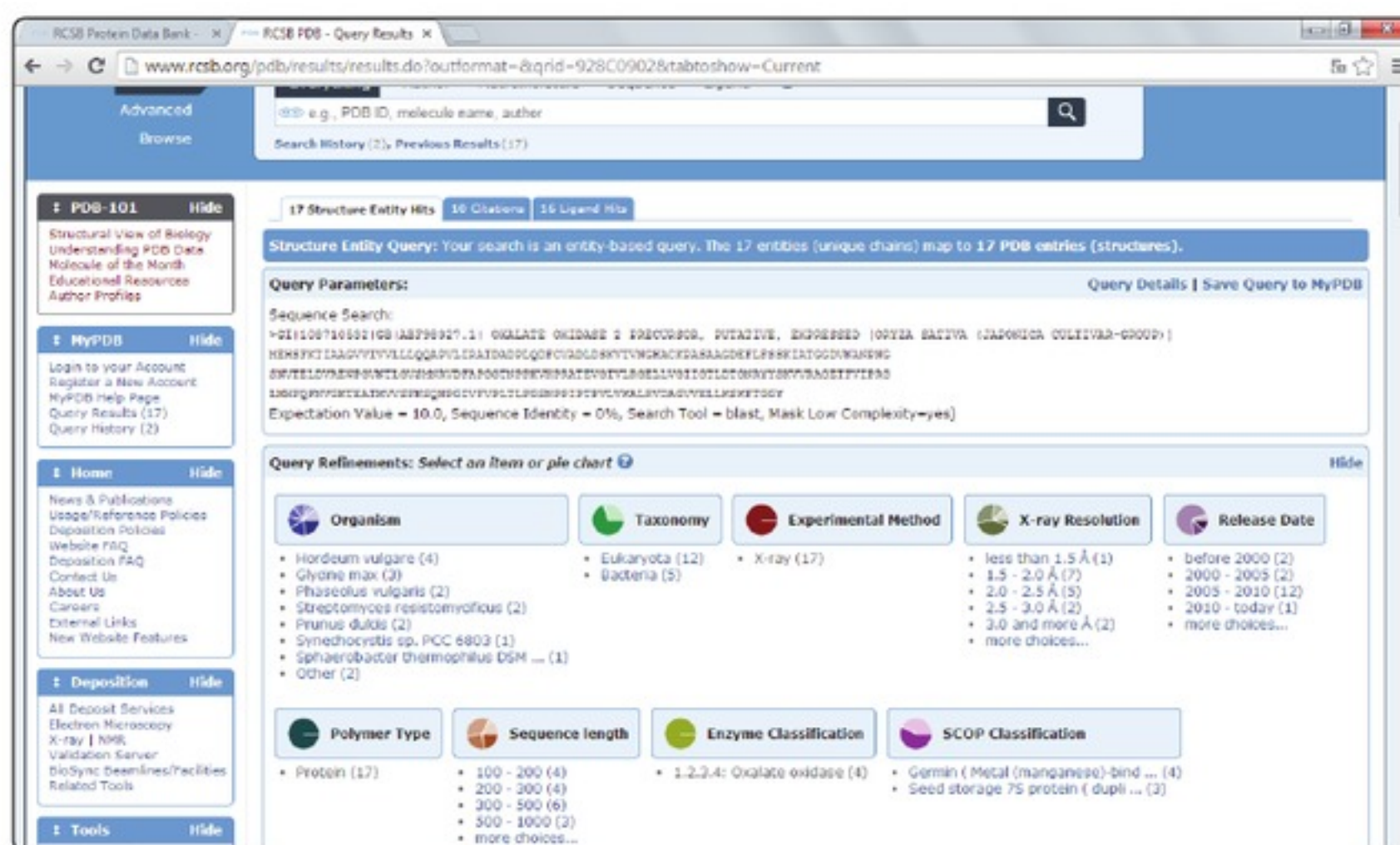


Figura 11.28. Resultados de la consulta en la base de datos de estructuras

Si pinchamos en el enlace correspondiente a una de las coincidencias en la base de datos, PDB muestra toda la información almacenada correspondiente a dicha coincidencia (ver Figura 11.29 y Figura 11.30):

The screenshot shows the PDB entry page for 1F12. The primary citation is: "Germin is a manganese containing homohexameric with oxalate oxidase and superoxide dismutase activities." by Woo, E.J., Dunwell, J.M., Goodenough, P.W., Marvier, A.C., Pickersgill, R.W. (2000) Nat. Struct. Biol. 7: 1036-1040. The molecular description indicates it is an Oxidoreductase, specifically Oxalate oxidase 1, with a structure weight of 21292.24. The 3D view shows a hexameric assembly of the protein.

Figura 11.29. Información de un registro PDB (I)

The screenshot shows the PDB entry page for 1F12, focusing on structure validation and experimental details. The structure validation section shows metrics like Clashscore, Ramachandran outliers, Sidechain outliers, and RSRZ outliers. The experimental details section shows the method (X-RAY DIFFRACTION), resolution (1.60 Å), and unit cell parameters.

Figura 11.30. Información sobre un registro PDB (y II)

En la sección **3D**, podemos ver una representación tridimensional de la estructura de la proteína construida con la herramienta Jmol, tal y como muestra la Figura 11.31:

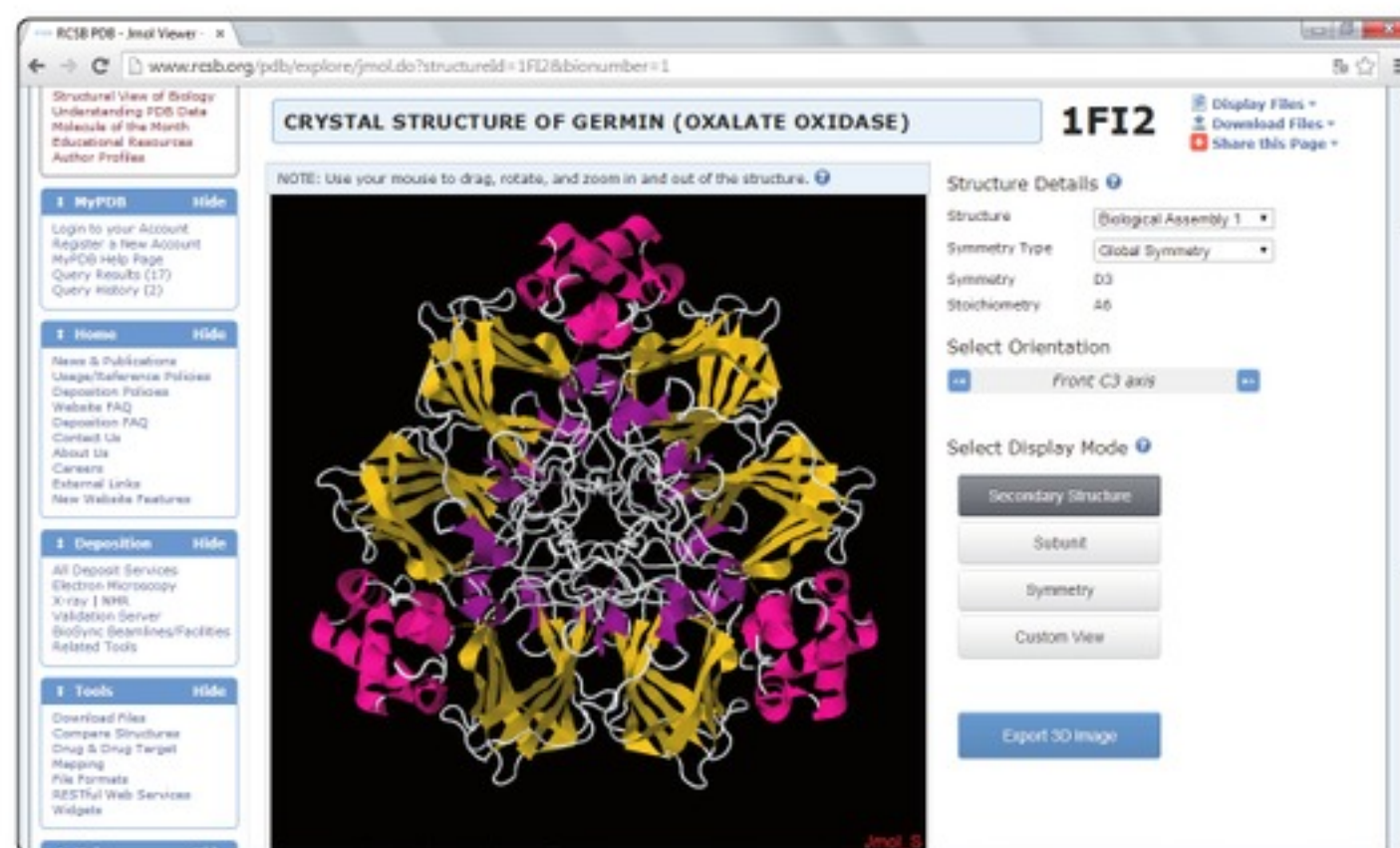


Figura 11.31. Estructura tridimensional de la proteína

La segunda pestaña de los resultados de la búsqueda son las citas académicas (ver Figura 11.32), donde obtenemos una lista de referencias que tratan sobre el resultado encontrado.

The screenshot shows the RCSB PDB Citation Results page. The title bar indicates the URL: www.rcsb.org/pdb/results/results.do?tab=show&Citation&qid=928C0902. The main header displays the search query: **OXALATE OXIDASE 2 PROTEIN, PUTATIVE, EXPRESSED (ORYZA SATIVA (JAPONICA CULTIVAR-BROOK))**. The page shows a list of search results with details for each entry, including the protein name, PDB ID, and a link to the full citation. The first result is for the protein 'OXALATE OXIDASE 2 PROTEIN, PUTATIVE, EXPRESSED (ORYZA SATIVA (JAPONICA CULTIVAR-BROOK))' with PDB ID 1FI2. The second result is for the protein 'OXALATE OXIDASE 2 PROTEIN, PUTATIVE, EXPRESSED (ORYZA SATIVA (JAPONICA CULTIVAR-BROOK))' with PDB ID 1FI2. The page also includes a 'Display/Download' button and a 'Sort by: 3 Publication Year' dropdown menu.

Figura 11.32. Resultados de la consulta en la base de datos de citas

Finalmente, la última pestaña es la de ligandos (ver Figura 11.33):

The screenshot shows the RCSB PDB Ligand Search results page. The search filters are set to 'Everything'. The search results display two ligands: ACETATE ION (ID: ACT, 2839 structures) and CALCIUM ION (ID: CA, 7589 structures). The page includes a search bar, navigation links, and a sidebar with PDB-101 and MyPDB sections.

Figura 11.33. Resultados de la búsqueda en la base de datos de ligandos

Cuando la proteína bajo estudio no tenga estructuras homólogas obvias, hay que recurrir a los algoritmos de predicción por reconocimiento de pliegues. Estos algoritmos tratan de alinear la estructura secundaria de la proteína bajo estudio con las almacenadas en una base de datos de estructuras



Modela la proteína bajo estudio lanzando una búsqueda con HHPred (<http://hhpred.tuebingen.mpg.de/hhpred>).

11.5 BÚSQUEDA DE MOTIVOS

Para encontrar motivos, utilizaremos las herramientas MEME y MAST, ambas disponibles en <http://meme.nbcr.net/meme/>.

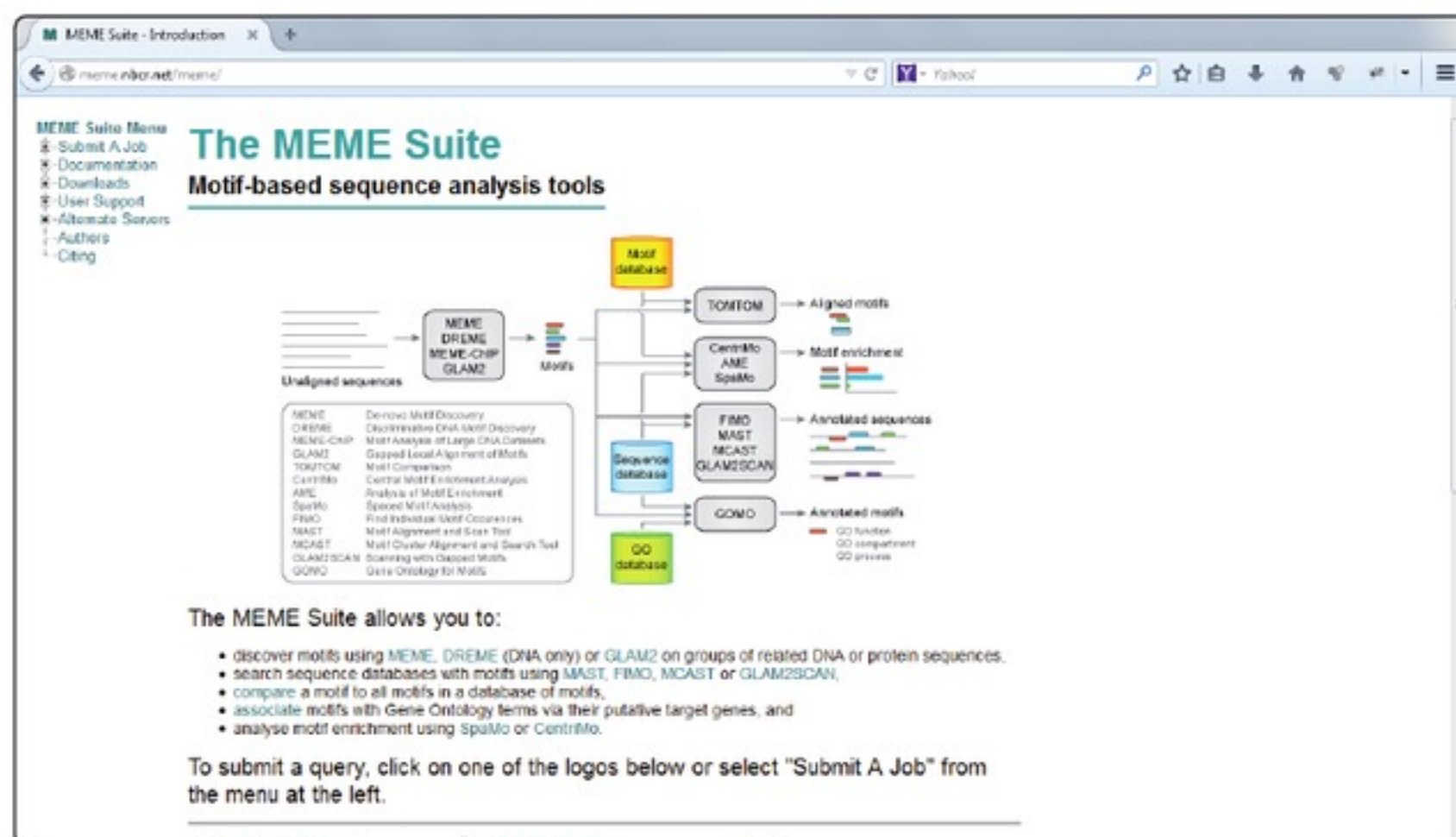


Figura 11.34. Página principal de MEME



Un motivo es un conjunto de bases conservadas que son importantes funcionalmente y que están localizadas próximas las unas de las otras.

Trabajaremos con las proteínas de ejemplo que pueden obtenerse pinchando en **Sample Protein Input Sequences**:

```
>At1g01140.1_4-2-4_SnRK3.12 SNF1-related Protein Kinase, subfamily 3
MSGSRKATPASRTRVGNYEMGRTLGECSFAKVYAKNTVTGDQAAIKILDREKVFRHKM
VEQLKREISTMKLIKHPNVVEIIIEVMASKTKIYIVLELVNGGELFDKIAQQGRLKEDEAR
RYFQQLINAVDYCHSRGVYHRDLKPENLILDANGVLKVSDFGLSAFSRQVREDGLLHTAC
GTPNYVAPEVLSDKGYDGAADVWSCGVILFVLMAGYLPFDEPNLMTLYKRICKAEFSCP
PWFSQGAKRVIKRILEPNPITRISIAELLEDEWFKKGYPSPFDQDDEDITIDVDAAFS
NSKECLVTEKKEKPVSMNAFELISSSSEFSLENLFQKQAQLVKKETRFTSQRSASEIMSK
MEETAKPLGFNVRKDNKYIKMKGDKSGRKGQLSVATEVFVAPSLHVVELRKTGGDTLEF
HKFYKNFSSGLKDVVWNTDAAAEEQKQ
>At1g01140.2_SnRK3.12 SNF1-related Protein Kinase, subfamily 3
MSGSRKATPASRTRVGNYEMGRTLGECSFAKVYAKNTVTGDQAAIKILDREKVFRHKM
```


VEQLKREISTMKLIKHPNVVEIIIEVMASKTKIYIVLELVNGGELFDKIAQQGRLKEDEAR
 RYFQQQLINAVDYCHSRGVYHRDLKPENLILDANGVLKVSDFGLSAFSRQVREDGLLHTAC
 GTPNYVAPEVLSDKGYDGAAADVWSCGVILFVLMAGYLPFDEPNLMTLYKRVRICKAEFS
 CPPWFSQGAKRVIKRILEPNPITRISIAELLEDEWFKKGYKPPSFDQDDEDITIDVDAA
 FSNSKECLVTEKKEKPVSMNAFELISSSSEFSLENLFKQAQLVKKETRFTSQRSASEIM
 SKMEETAKPLGFNVRKDNYKIKMKGDKSGRKGQLSVATEVFVAPSLHVVELRKTGGDTL
 EFHKFYKNFSSGLKD VVWNTDAAAEEQKQ

>At1g01140.3_SnRK3.12 SNF1-related Protein Kinase, subfamily 3
 MSGSRRKATPASRTRVGNYEMGRTLGECSFAKVYAKNTVTGDQAAIKILDREKVFRHKM
 VEQLKREISTMKLIKHPNVVEIIIEVMASKTKIYIVLELVNGGELFDKIAQQGRLKEDEAR
 RYFQQQLINAVDYCHSRGVYHRDLKPENLILDANGVLKVSDFGLSAFSRQVREDGLLHTAC
 GTPNYVAPEVLSDKGYDGAAADVWSCGVILFVLMAGYLPFDEPNLMTLYKRICKAEFSCP
 PWFSQGAKRVIKRILEPNPITRISIAELLEDEWFKKGYKPPSFDQDDEDITIDVDAAFS
 NSKECLVTEKKEKPVSMNAFELISSSSEFSLENLFKQAQLVKKETRFTSQRSASEIMSK
 MEETAKPLGFNVRKDNYKIKMKGDKSGRKGQLSVATEVFVAPSLHVVELRKTGGDTLEF
 HKVCDSFYKNFSSGLKD VVWNTDAAAEEQKQ

>At1g01450.1_2-1-1 putative protein kinase
 MADFLLKHLGDGNESPKLFPSSLLDNTKDYQVKKRLNGSQYKEITWLGESFALRHFFGD
 IDALLPQITPLLSLHPNIVYYLCGFTDEEKKECFLVMELMRKTLGMHIKEVCGPRKKNT
 LSLPVAVDLMLQIALGMEYLHSHKRIYHGELNPSNVLKPRSNQSGDGYLLGKIFGFGLNS
 VKGFSSKSASLTSQENENFPFIWYSPEVLEEQEQSGTAGSLKYSKSDVYSFGMVSEFLLT
 GKVPFEDSHLQGDKMSRNIRAGERPLFPFNSPKFITNLTKRCWHADPNQRPTFSSISRIL
 RYIKRFLALNPECYSSSQDPSIAPTVDYCEIETKLLQKLSWESTELTKVSQVPFQMFAY
 RVVERAKTCEKDNLREPSESGSEWASCEDEGGAGSDEQLSYAKERRLSCSSNDVGMSKK
 QVSNLLKRASSLKPIQKPGEEIIISQYIYIYIGSLTNMNLVTCTNFFVLCH

>At1g01540.1_1-6-3 Putative protein kinase
 MSVYDAAFLNTELSKPTSIFGLRLWVIGILLGSLIVIALFLLSLCLTSRRKNRKPRADF
 ASAAIATPPISKEIKEIVPAQNQSVPAEIQVDIGKIEHRVVFSDRVSSGESRGTSASSET
 ASYSGSGNCGPEVSHLGWGRWYTLRELEAATNGLCEENVIGEGGYGIVYRGILTDGTKVA
 VKNLLNNRGQAEKEFKVEVEVIGRVRHKNLVRLGVCVEGAYRMLVYDFVDNGNLEQWIH
 GDVGDVSPLTWDIRMNIILGMAKGLAYLHEGLEPKVVHRDIKSSNILLDRQWNAKVSDFG
 LAKLLGSESSYVTTTRVMGTFGYVAPEYACTGMLNEKSDIYSFGILIMEIITGRNPVDYSR
 PQGEVFDKHIQSSLCFCKWSYVSWL

>At1g01540.2_Putative protein kinase
 MSVYDAAFLNTELSKPTSIFGLRLWVIGILLGSLIVIALFLLSLCLTSRRKNRKPRADF
 ASAAIATPPISKEIKEIVPAQNQSVPAEIQVDIGKIEHRVVFSDRVSSGESRGTSASSET
 ASYSGSGNCGPEVSHLGWGRWYTLRELEAATNGLCEENVIGEGGYGIVYRGILTDGTKVA
 VKNLLNNRGQAEKEFKVEVEVIGRVRHKNLVRLGVCVEGAYRMLVYDFVDNGNLEQWIH
 GDVGDVSPLTWDIRMNIILGMAKGLAYLHEGLEPKVVHRDIKSSNILLDRQWNAKVSDFG
 LAKLLGSESSYVTTTRVMGTFGYVAPEYACTGMLNEKSDIYSFGILIMEIITGRNPVDYSR
 PQGETNLVDWLKSMVGNRRSEEVDPKIPPEPSSKALKRVLLVALRCVDPDANKRPMKGH
 IIHMLEAEDLLYRDERRTTRDHGSRERQETAVVAAGSESGESGRHHQQKQR

>At1g01560.1_4-5-1_MPK11 MAP kinase 11
 MSIEKPPFGDDSNRGVSINGGRYVQYNYGNLFVSKKYVPPLRPIGRGASGIVCAAWNS
 ETGEEVAIKKIGNAFGNIIDAKRTLREIKLLKHMDHDNVIAIIDIIIRPPQPDNFNDVHIV

YELMDTDLHHIIRSNQPLTDDHSRFFLYQLLRGLKYVHSANVLHRDLKPSNLLLNANCDL
KIGDFGLARTKSETDFMTEYVVTRWYRAPELLNCSEYTA AIDIWSVGCILGEIMTREPL
FPGRDYVQQRLRLITEVNFSLSFHLTILFRFNLKKEH

>At1g01740.1_1-16-1 putative protein kinase

MGGQSSKIGTCCSHKTTALEAPDVENKENGEVNGVHSFREYSLEQLKIATSCFALENVVS
EHGETAPNVVYQGKLENHMKIAIKRFSGTAWPDPRQFLEEARLVGQLRSKRMANLLGYCC
EGGERLLVAEFMPNETLAKHLFHWDTEPMKWAMRLRVALYISEALEYCSNNGHTLYHDLN
AYRVLFDEECNPRNSTFGLMKNSRDGKSYSTNLAFTPEYLRTGRITAESVIYSFGTLLL
DLLTGKHIPPSHALDLIRDRLNQLTLDSCLEGQFSDSDGTELVRLTSCCLQYEARERPNI
KSLVTALISLQKDTEVLSHVLMGLPQSGTFASPPSPFAEACSGKDLTSMVEILEKIGYKD
DEDLSFMWTEQMQEAINSKKKGDIAFRRKDFSEAIEFYTQFLDLGMISATVLVRRSQSYL
MSNMAKEALDDAMKAQGISPVWYVALYLQSAALSVLGMEKESQIALTEGSILEARKISAS
TQN

>At1g02970.1_4-3-1 putative protein kinase

MFEKNGRTLLAKRKTQGTIKTRASKKIRKMEGTLEHSLQFGQLSKISFENRPSSNVAS
SAFQGLLDSDSSELRLNQLGSADSDANCGEKDFILSQDFFCTPDYITPDNQNLMGSLDISK
DHSPCPRSPVKLNTVKSKRCRQESFTGNHSNSTWSSKHRVDEQENDDIDTDEVMGDKLQA
NQTERTGYVSQAVALRCRAMPPCLKNPYVLNQSETATDPFGHQRSKASFLPVSTSGD
GLSRYLTDFHEIRQIGAGHFSRVFKVLKRMDCGLYAVKHSTRKLYLDSERRKAMMEVQAL
AALGFHENIVGYSSWFENEQLYIQLELCDHSLSALPKKSSLKVSEREILVIMHQIAKAL
HFVHEKGIAHLDVKPDNIYIKNGVCKLGDGFCATRLDKSLPVEEGDARYMPQEILNEDYE
HLDKVDIFSLGVTYELIKGSPLTESRNQSLNIKEGKLPLLPGHSLQLQQLLKTMMDRDP
KRRPSARELLDHPMFDRIRG

>At1g03740.1_4-5-2 putative protein kinase

MGCVNSRHRPFRRKSTTLKESSEEKRSSRIDSSRRIDDWIQPEDGFDRLSNSGDAKVRLI
ESEMFTSRCHDHQIGKILENPATVAHMDRVVHDQELRRASSAVVDSDLIDPKVVKAKL
DRWNSKDSKVRLIESEKLSSSMFSEHHQIEKGVEKPEVEASVRVHRELKRGSSIVSPKD
AERKQVAAGWPSWLVS VAGESLVDWAPRRANTFEKLEKIGQGTYSVYRARDLLHNKIVA
LKKVRFDLNDMESVKFMAREIIVMRRLDHPNVLKLEGLITAPVSSSLYLVFEYMDHDLG
LSSLPGVKFTEPQVKCYMRQLLSGLEHCHSRGVLHRDIKGSNLLIDSKGVLKIADFGLAT
FFDPAKSVSLTSHVVT LWYRPPELLLGASHYGVGVDLWSTGCILGELYAGKPILPGKTEV
EQLHKIFKLCGSPTENYWRKQKLPSSAGFKTAIPYRRKVSEMFKDFPASVLSLLETLLSI
DPDHRSSADRALESEYFKTKPFACDPSNLPKYPPSKEIDAKMRDEAKRQQPMRAEKQEDK
TL

>At1g03920.1_4-2-6 putative protein kinase

MDSARSWFHKFQPRDKPRKKDMFSGSTYGGGVTTETVDPDGGNDTETATKLPPLGGDGEAL
SNSTKQKVAAAKQYIENHYKEQMKNLNERKERRTTLEKKLADADVCEEDQTNLMKFLEKK
ETEYMLRQRHKMGADDFELLTMIGKGAFGEVRVREINTGHVFAMKKLKKSEMLRRGQVE
HVRAERNLLAEVDSNCIVKLYCSFQDNEYLYLIMEYLPGGDMMTLLMRKDTLSEDEAKFY
IAESVLAIESIHNRYIHRDIKPDNLLLDYRGLRLSDFGLCKPLDCSVIDGEDFTVGNA
GSGGGSESSTTPKRSQQEQLEHWQKNRRMLAYSTVGTDPDYIAPEVLLKKGYGMECDWWS
LGAIMYEMLVGYPPFYADDPMSCTRKIVNWKTHLKFPEESRLSRGARDLIGKLLCSVNQR
LGSTGASQIKAHWPFEQVQWEKIYQMEAAFIPEVNDDLDTQNFEKFDEEDNQTPAPSRTG
PWRKMLSSKDINFVGYTYKNFEIVNDYQVPGIAELKKKESKSKRPSVKSLFESESDSSSS
GSEQQTINRSYSNPTPRGMEPNLRRLDSE


```
>At1g03930.1_3-1-1-1_ADK1 protein kinase ADK1
MDLVIGGKFKLGRKIGSGSFGELYLGINVQTGEEVAVKLESVKT KHPQLHYESKLYMLLQ
GGTGVPNLKWYGVEGDYVMVIDLLGPSLEDLFNYCNRKLSLKTVMMLADQLINRVEFMH
TRGFLHRDIKPDNFLMGLGRKANQVYIIDFGLGKKYRDLQTHRHI PYRENKNLTGTARYA
SVNTHLGVEQSRDDLEALGYVLMYFLKGSLPWQGLKAGTKKQKYDRISEKKVATPIEVL
CKNQPSSEFVSFVRYCRSLRFDDKPDYSYLRDLFRDLFIREGYQFDYVFDWTVLKYPQIGS
SSGSSSRTRNHTTANPGLTAGASLEKQERIAGKETRENRFSGAVEAFSRRHPATSTTRDR
SASRNSVDGPLSKHPPGDSEPRSSSRYSRRRAIPSSSRPSSAGGPSDSRSSSRRLVTS
TGGVGTVSNRASTSQRISQAGNESRTSSFSRAARNTREDPLRRSLELLTLRK
```

```
>At1g04210.1_protein kinase ADK1
MDSKIKKPANLIEDADIDGGSESDSTISSVLSLEDDSVVDVSGQNLEFSLLDNVDDSVKG
LYFFRNVFNLIPKSIIGGLGRRLKLFSSNEIDLFPPELGNLVNLEYLQVKISSPGFGDGL
SWDKLKGLKELELTKVPRSSALTLLSEISGLKCLTRLSVCHFSIRYLPPEIGCLKSLEY
LDLSFNKIKSLPNEIGYLSSLTFLKVAHNRLMELSPVLALLQNLES LDVSNNRLTTLHPL
DLNLMPLRLQILNLRYNKLPSYCWIPWIIQCNFEGNYEEMGVDTCS SSMVEMDVFETPYEN
NVITVPHKGS HRNPLNMSTGISSISRCFSARKSSKRWRKRRQYYFQQRARQERLNN SRKWK
GEVPPEGLSLKMEVEETGKQGMKVPQNTDRGSVDNSCSDENDKLFEEASVITSEEEESSL
KADVVDNSQCVETQLTSERDNYESCEIKTSSPSSGDAPGTVDYNSSSERKKPNKSKRC
SEKYLDNPKGSKCHKLSTDITNLSRKYSNSFCSTEDSLPDGFFDAGRDRPFMTLSKYEK
VLPLDSREVILLDRAKDEVLDAITLSARALVARLKKLNCLTPDVDQVSIDNLQVASFLAL
FVSDHFGGSDRTAI IERTRKAVSGTNYQKPFICTCLTGNQDDLAALNKQVSTTAEDAILS
DVCEKSLRSIKSKRNSIVVPLGKLQFGICRHRALLMKYLCDRMEPPVPCELVRGYLDFMP
HAWNIVPVKQGSSWVRMVVDACRPHDIREDTDQEYFCRYIPLNRLNESIRIKEKLEPGCS
VSSLSTGKGV ERANSSLIRCKLGSTEAVVKMRTLEVSGASLDDIRTFEYTCLGEVRILGA
LKHDCIVELYGHEISSKWITSENGNEHRVLQSSILMEHIKGGSLKGHIEKLSEAGKHHVP
MDLALSIARDISGALMELHSDIHRDIKSENVLIDLNDQSANGEPIVKLCDFDRAVPLR
SHLHGCCIAHVGI PPPNICVGTPRWMSPEVFRAMHEQNFGYGLEVDIWSFGCLIFELLTLQ
NPYFDLSELQIHESLQNGKRPKLPKKLET LISETEEEEESTNKLSEVFDLTESDLDTMRFL
IDVFHQCTEESPSDRLNAGDLHEMILSRKKRE
```

```
>At1g04440.1_3-1-1-1 putative casein kinase I
MDRVVGGKFKLGRKLGSFGSFGEIFLG VNVQTGEEVAVKLEPLRARHPQLHYESKLYMLLQ
GGTGIPHLKWFGVEGEFNCMVIDLLGPSMEEFFNYCSRSFSLKTVMMLADQMINRVEYMH
VKGFLHRDIKPDNFLMGLGRKANQVYIIDYGLAKKYRDLQTHKHIPYRENKNLTGTARYA
SVNTHLGIEQSRDDLES LGYLLMYFLRGSLPWQGLRAGTKKQKYDKISEKKRLTPVEVL
CKNFPPEFTSYFLYVRSLRFEDKPDYSYLRDLFRDLFIREGYQFDYVFDWTILRYPQFGS
SSSSNSKPRPTLRPAMNIPVPSADKA EKPPIGQDSRERFSGVFEAYTRRNGSGTGVQADQ
SSRPRTSENVLASKDTQNQERPNSLSRNLSSSRKAIAGSSVRATSSADFTENRLSRLIPN
NDRSSTTLRTQFAPSSSSSVATKAAPTRAARDITLQSLLELLSIGNSKRK
```

```
>At1g04700.1_2-1-4-1_Raf16 MAP kinase kinase kinase Raf16
MRMEFPGSSNQHLGRDRFNGEVGCNNCSQTGEEFSNEFLRDFGAQRRLQHGGVNRNVEG
NYYNRHLVYEDFNRLGLQRVDSNMSEGINSSNGYFAESNVADSPRKMFTAISDVYLPE
VLKLLCSFGGRILQRPDGDGLRYIGGETRIISIRKHVGLNELMHKTYALCNHPHTIKYQL
PGEDLDALISVCSDEDLLHMIEEYQEAETKAGSQRIRVFLVPSTESSESPKIFHERNMNI
NRNTNQQT DIDHYQYVSALNGIVDVSPQKSSSGQSGTSQTTQFGNASEFSPTFHLRDSPT
SVHTWEHKDSNSPTFMKPYGNTNAVHFMPKMQIPRNSFGQQSPPTSPFSVHKRANTDVPY
```


FADQNGFFDPYLAAPNFPQQNRFFFETTTQKQKHPEVNLHDRRPSDDIYPHGQAYIGA EK
 MTLKKNALSDPQLHDESQINNGLEAFTKQPWKILRKNLRVVATSKWEDSDDIYFNNPEGK
 RCKELELTKEVPNSWINRDNNPDSFDQATKKQDGSNSNSSFSFSPNYFSPNHQPAAQITSSD
 SQDSGSSVFSLSVNTNENYLDSCREKFNGFQHDMSLDILIRSHTSATDQLCSTTKSSDKA
 DYSSPNTNFPVVFLRQEPMI PRHDLETNSDDSDTQKSLPREESIHYSGLPLRKVGSRETT
 FMHTQGSDDFFKSKLLGPQLIVEDVTNEVISDNLLSATIVPQVNRESDDDHKSYTREKEI
 TNADHESEMEEKYKKS RNTDDSFSEAAMVEIEAGIYGLQIIKNTDLEDLHELGS GTFGTV
 YYGKWRGTDVAIKRIKNSCFSGGSSEQARQTKDFWREARILANLHHPNVVAFYGVVPDGP
 GGTMATVTEYMVNGSLRHVLQRKDRLLDRRKKLMITLDSAFGMEYLHMKNIVHFDLKCDN
 LLVNL RDPQRPICKVGDFGLSRIK RNTLVSGGVRGTLPWMAPELLNGSSNRVSEKVDVFS
 FGIVMWEILTGE EPYANLHCGAIIGGIVNNTLRPPVPERCEAEWRKLMEQCWSFDPGVRP
 SFTEIVERLRSM TVALQPKRRT

>At1g05100.1_4-4-1_MAPKKK18 MAP kinase kinase kinase 18

MNWSTRGKTLGRGSTATVSAATCHESGETLAVKSAEFHRSEFLQREAKILSSLNSPYVIGY
 RGCEITREPFHNNGEATTYSLLMEYAPYGTLT DVATKNGGFIDEARVVKYTRQILLGLE Y
 IHNSKGIAHCDIKGSNVLVGENGEAKIADFGCAKWVEPEITEPVRGTPAFMAPEAARGER
 QGKESDIWAVGCTVIEMVTGSQPWIGADFTDPVSVLYRVGYLGELPELPCSLTEQAKDFL
 GKCLKKEATERWTASQLLNHPFLVNKEPELVTGLVTNSPTSVDQMFWRSV EEEVSEDRS
 SWWECHEDERIGVLSWIGHVVVESTWDLDGEDWITVRN

>At1g05700.1_1-8-1 putative light repressible receptor protein

MEEFRFLYLIYSAAFALCLVSVLAQDQSGFISIDCGIPSGSSYKDDTTGINYVSDSSSFV
 ETGVSKSIPFTAQRQLQNLR SFPEGSRNCYTLIPIQGKGKKYLIRASFMYGNYDGENGSP
 EFDLFLGGNIWDTVLLSNGSSIVSKEVVYLSQSENIFVCLGNKGKGT PFI STLELRLFLGN
 DNTTYDSPNGALFFSRRWDLRSLMGSPVRYDDD VYDRIWIPRNFGYCREINTSLPVTSDN
 NSYSLSSLVMSTAMTPINTTRPITMTLENSDPNVRYFVYMHFAEVEDLSLKPNQTREFDI
 SINGVTVAAGFSPKYLQTN TFFLN PESQSKIAFSLVRTPKSTLPPIVNALEIYVANSFSQ
 SLTNQEDGDAVTS LKTSYKVKNWHGDPCLPNDYIWEGLNCSYDSLTPPRITSLNLSSSG
 LTGHIS SFSNL TMIQELDLSNNGLTGDIPEFLSKLKLFLRVLNLENNTLTG SVPSELLER
 SNTGSFSLRLGENPGLCTEISCRKSNSKKLVIPLVASFAALFILLLLSGVFWRIRNRN
 PMAKSENKLLFTFADVIKMTN NFGQVLGKGGFGTVYHGFYDNLQVAVKLLSETSAQGFKE
 FRSEVEVLVRVHHVNLTALIGYFHEGDQMG LIYEFMANGNMADHLAGKYQHTLSWRQRLQ
 IALDAAQGLEYLHCGCKPPIVHRDVKTSNILLNEKNRAKLADFGLSRSFHTESRSHVSTL
 VAGTPGYLDPLCFETNGLNEKSDIYSFGVV LLEMITGKTVIKESQTKRVHVSDWVISILR
 STNDVNNVIDSKMAKDFDVNSVWKVVELALSSVSQNVSDRPNMPHIVRGLNECLQREESN
 KNY

>At1g06390.1_4-5-4_ASK-iota GSK3/shaggy-like protein kinase iota

MASLPLGPQPHALAPPLQLHDGDALKRRPELDSDKEMSAAVIEGNDAVTGHIISTTIGGK
 NGEPKQTISYMAERVVGTGSFGIVFQAKCLETGESVAIKKVLQDRRYKNRELQLMRPMDH
 PNVISLKHCFSTTSRDEFLNLVMEYVPETLYRVLRHYTSSNQRMPIFYVKLYTYQIFR
 GLAYIHTVPGVCHRDVKPQNLLVDPLTHQVKLCDFGS AKVLVKGEPNISIYICSRYYRAPE
 LIFGATEY TASIDIWSAGCVLAELLGQPLFPGENSVDQLVEIIKVLGTP TREEIRCMNP
 NYTDFRFPQIK AHPWHKVFHHRMPPEAIDLASRL LQYSPSLRCTALEACAH PFFNELREP
 NARLPNGRPLPPLFNFKQELGGASME LINRLIPEHVR RQMSTGLQNS

>At1g06390.2_ASK-iota GSK3/shaggy-like protein kinase iota

MASLPLGPQPHALAPPLQLHDGDALKRRPELDSDKEMSAAVIEGNDAVTGHIISTTIGGK


```

NGEPKQTISYMAERVVGTGSFGIVFQAKCLETGESVAIKKVLQDRRYKNRELQLMRPMDH
PNVISLKHCFSTTSRDELFNLVMEYVPETLYRVLRHYTSSNQRMPIFYVKLYTYQIFR
GLAYIHTVPGVCHRDVQPNLLVDPLTHQVKLCDFGSAKVLVKGEPNISIYICSRYYRAPE
LIFGATEYASIDIWSAGCVLAELLGQPLFPGENSVQDLVEIIKVLGTPTREEIRCMNP
NYTDFRFPQIKAHPWHKVFHHRMPPEAIDLASRLQYSPSLRCTALEACAHPPFNELREP
NARLPNGRPLPPLFNFKQELGGASMELINRLIPEHVRQMQSTGLQNS

```

Figura 11.35. Proteínas bajo estudio

Una vez en la página principal de la *suite* de herramientas MEME, pinchamos en el icono correspondiente a MEME situado en la parte izquierda de la pantalla y se mostrará el formulario de introducción de datos de búsqueda (ver Figura 11.36):

MEME Suite Menu

- Submit A Job
- Documentation
- Downloads
- User Support
- Alternate Servers
- Authors
- Citing

MEME
Multiple Em for Motif Elicitation
Version 4.9.1

Use this form to submit DNA or protein sequences to MEME. MEME will analyze your sequences for similarities among them and produce a description (**motif**) for each pattern it discovers.

Data Submission Form

Required

Your e-mail address:

Re-enter e-mail address:

Please enter the **sequences** which you believe share one or more motifs. The sequences may contain no more than **60000 characters** total total in any of a large number of **formats**.

Enter the name of a file containing the sequences here:
 No se ha seleccionado ningún archivo.

or
 the **actual sequences** here (Sample Protein Input Sequences):

How do you think the occurrences of a single motif are **distributed** among the sequences?

☐ One per sequence
☒ Zero or one per sequence
☐ Any number of repetitions

MEME will find the optimum **width** of each motif within the limits you specify here:

Minimum width (>= 2)
 Maximum width (<= 300)
 Maximum number of motifs to find

Options

Description of your sequences:

MEME will find the optimum **number of sites** for each motif within the limits you specify here:

Perform **discriminative** motif discovery – Enter the name of a file containing **'negative sequences'**:
 No se ha seleccionado ningún archivo.

Figura 11.36. Parámetros de búsqueda de MEME

En primer lugar, arriba a la izquierda, se pide la información necesaria para enviar los resultados a una dirección de correo electrónico. También arriba, pero a la derecha, se establece el número de motivos que se espera encontrar en la secuencia, así como la longitud estimada de dichos motivos.

Por otra parte, la opción **Shuffle sequence letters** indica si deseamos barajar las letras de la secuencia de entrada. Esto resulta bastante útil para determinar si los motivos encontrados con la sin barajar son estadísticamente significativos o no. Para ello, se compara el valor E del mejor motivo/alineamiento de la secuencia original con el correspondiente calculado con la opción **Shuffle sequence letters** marcada. Si son iguales, probablemente el motivo sea poco significativo. También es posible indicar si queremos realizar la búsqueda únicamente en la hebra proporcionada (**Search given strand only**) y/o si se desea buscar secuencias palindrómicas (**Look for palindromes only**).

Finalmente, disponemos de un área en la que introducir la secuencia proteica o bien su número de *accession*. En nuestro caso, pegaremos la secuencia de aminoácidos en formato FASTA y pulsaremos el botón **Start search** para lanzar la búsqueda y obtendremos una pantalla intermedia como la de la Figura 11.37, en donde se nos proporciona el enlace a los resultados del proceso junto con un resumen de la información de configuración de la búsqueda que hemos introducido anteriormente.

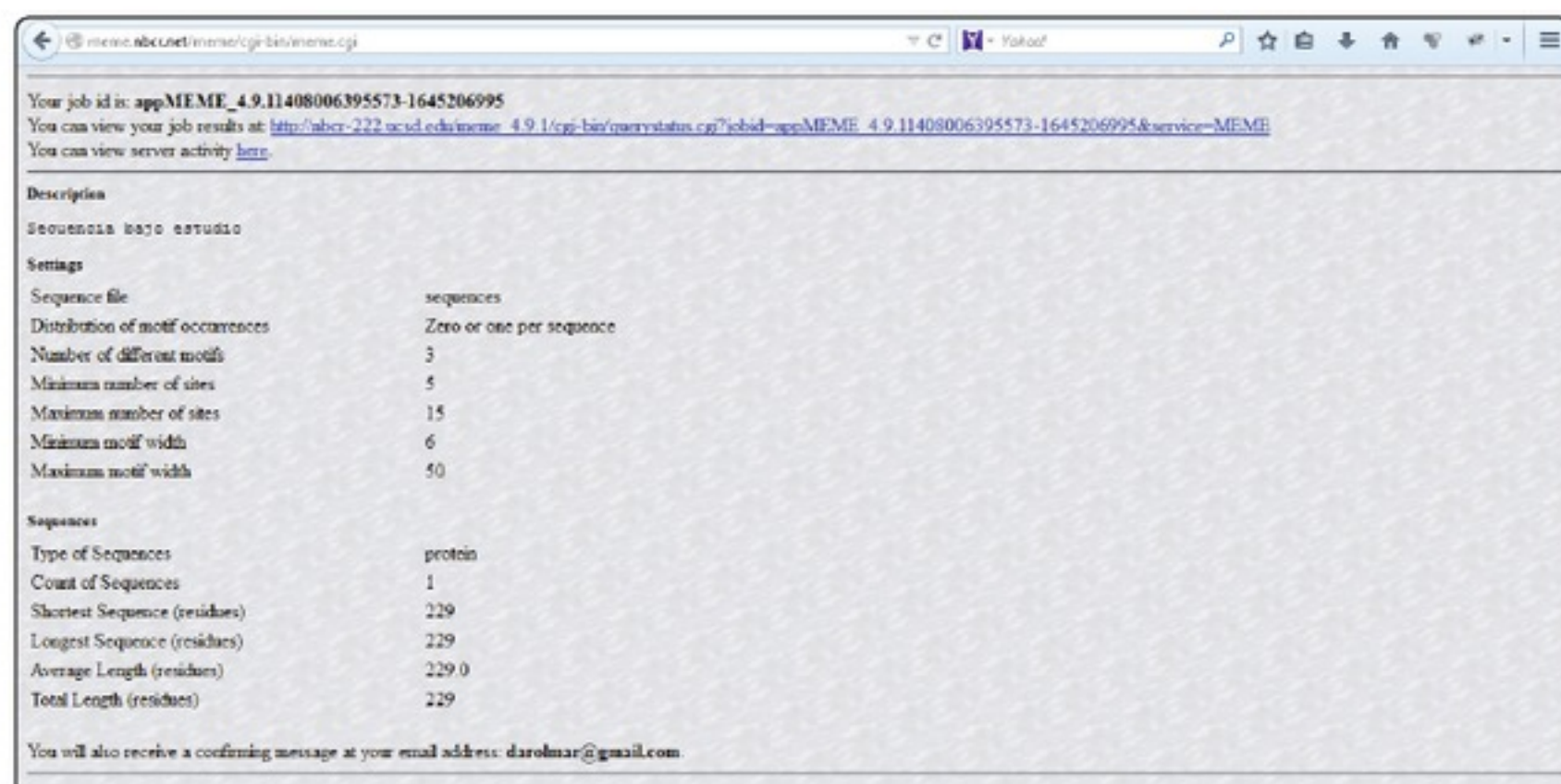


Figura 11.37. Página de resultados de MEME

Pinchando en el enlace **You can view your job results at**, se muestra una página en la que podemos consultar el estado del trabajo. Cuando este esté listo, aparecerá una pantalla como la de la Figura 11.38:



Figura 11.38. Resultado final de MEME

Pulsando, por ejemplo, en **MEME.html output**, obtenemos los resultados en formato HTML (ver Figura 11.39). Esta página, de extensión considerable, se divide en varias secciones.

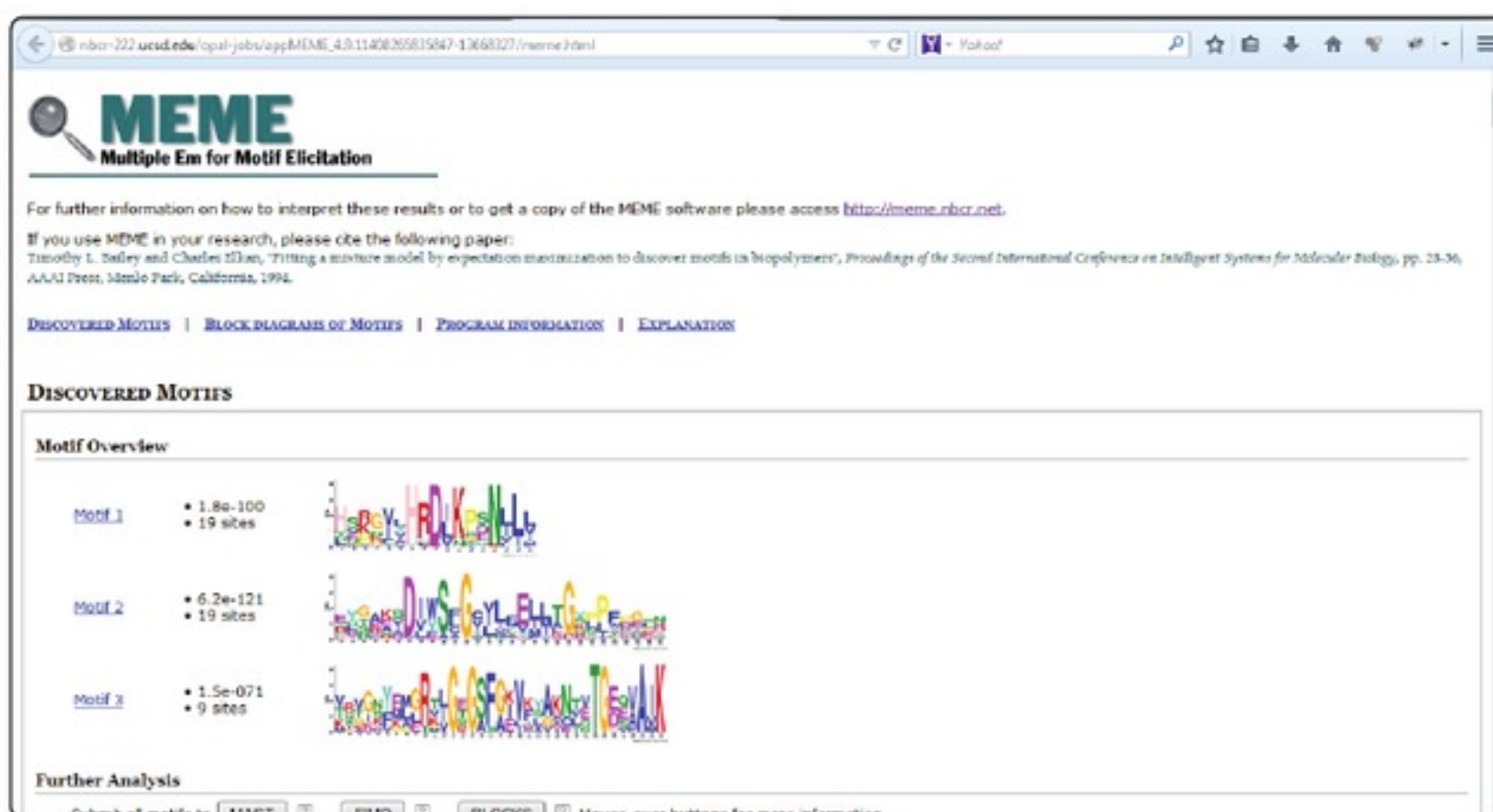


Figura 11.39. Página de resultados de MEME en formato HTML

En la sección de motivos encontrados, obtenemos información sobre los motivos hallados, en este caso, tres. Para cada motivo se ofrece su presentación en forma de logo y un enlace que nos lleva a una descripción detallada de dicho motivo. Si pinchamos, por ejemplo, el correspondiente al primer motivo, *Motif1*, encontramos la información organizada en varias partes.

En primer lugar, encontramos un resumen de la información sobre el motivo y el logo de su secuencia (ver Figura 11.40). En el resumen se ofrece la siguiente información:

- Valor E (*E-value*): importancia estadística del motivo. MEME muestra los motivos con mayor importancia estadística (menor valor E) los primeros en la lista. El cálculo del valor E de un motivo se basa en su ratio de probabilidad, amplitud, sitios, frecuencias de las letras de fondo y el tamaño del conjunto de entrenamiento y constituye una estimación del número de motivos esperado dado un ratio de probabilidad y con la misma amplitud y número de sitios que uno encontraría de un conjunto de secuencias aleatorias de tamaño parecido.
- Amplitud (*Width*): amplitud del motivo. Cada motivo describe un patrón de una amplitud fija, puesto que MEME no permite huecos.
- Sitios (*Sites*): número de sitios implicados en la construcción del motivo.
- Ratio de probabilidad (*Log Likelihood Ratio*): ratio de probabilidad del motivo, en unidades logarítmicas. Mide la probabilidad de ocurrencias del motivo, dado un modelo del motivo frente a la probabilidad de ocurrencia del motivo sin disponer de modelo del mismo.
- Contenido de la información (*Information Content*): es el motivo en bits.
- Entropía relativa (*Relative Entropy*): es la entropía relativa del motivo en bits y se calcula como el ratio de probabilidad dividido entre el número de *sites*.

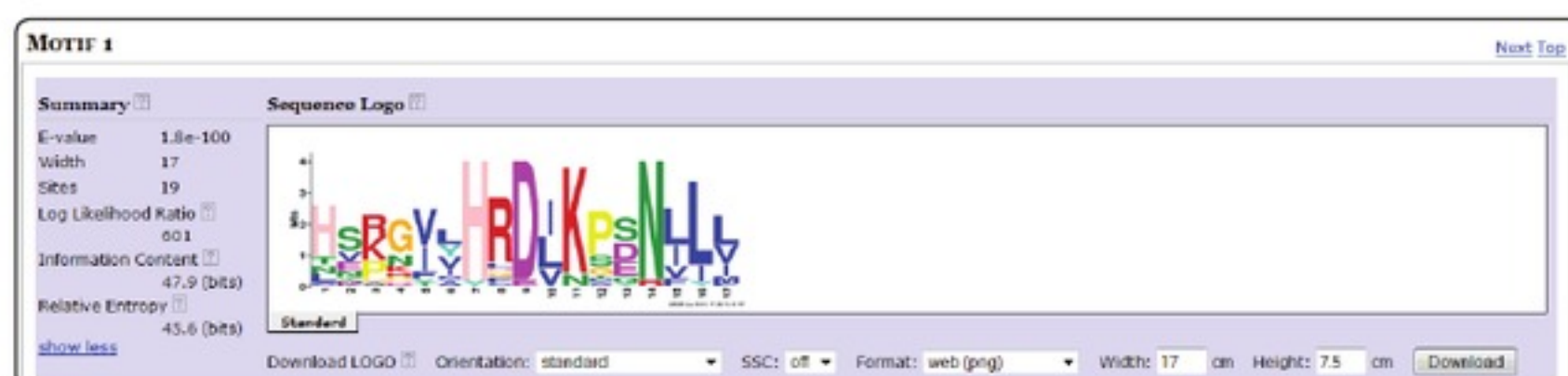


Figura 11.40. Resumen y logo del motivo Motif1

A continuación, encontramos una expresión regular que describe el motivo, si queremos efectuar un análisis más profundo del motivo en cuestión (entraremos en detalle más adelante en este epígrafe) y de los formatos de salida (ver Figura 11.41):

Regular expression

Further Analysis Mouse-over buttons for more information.

Data Formats or

Figura 11.41. Información adicional del motivo

Seguidamente, se muestra la información sobre los sitios (ver Figura 11.42). MEME muestra los sitios (ocurrencias) del motivo en el conjunto de entrenamiento. Estos sitios se alinean entre sí y las diez posiciones anteriores y posteriores a cada sitio también se visualizan. Cada sitio se identifica por el nombre de la secuencia en que ocurre, la habrá si se emplean las dos hebras de la secuencia de ADN y la posición en la secuencia en la que el sitio comienza.

Sites

Name	Start	p-value	Sites
At1g01140.3_SnRK3.12	100	3.47e-25	V A P E V L S D K G Y D G A A A D V W S C G V I L F V I M A G Y L F D E F N L M T L Y K R I C K
At1g01140.2_SnRK3.12	100	3.47e-25	V A P E V L S D K G Y D G A A A D V W S C G V I L F V I M A G Y L F D E F N L M T L Y K R V R I
At1g01140.1_4-2-4_SnRK3.12	196	3.47e-25	V A P E V L S D K G Y D G A A A D V W S C G V I L F V I M A G Y L F D E F N L M T L Y K R I C K
At1g06390.2_ASK-iota	247	1.06e-24	R A P E L I F G A T E Y T A S I D I W S A G C V L A E L L L Q Q F L P G E N S V D Q L V E I I K
At1g06390.1_4-5-4_ASK-iota	247	1.06e-24	R A P E L I F G A T E Y T A S I D I W S A G C V L A E L L L Q Q F L P G E N S V D Q L V E I I K
At1g01560.1_4-5-1_MPK11	217	2.12e-23	R A P E L L I N C S E Y T A A I D I W S V G C I L G E I M T R E L F P G R D Y V Q Q L R L I T E
At1g01540.2_Put30ve	332	3.16e-21	Y V A P E Y A C T G M L N E K S D I Y S F G I L I M E I I T G R N F V D Y S R P Q G E T N L V D W
At1g01540.1_1-6-3	332	3.16e-21	Y V A P E Y A C T G M L N E K S D I Y S F G I L I M E I I T G R N F V D Y S R P Q G E V F D K H I
At1g03920.1_4-2-6	351	3.73e-20	Y I A P E V L L K K G Y G M E C D W S L G A I M Y E M L V G Y P P F A D D P M S T C R K I V N
At1g01450.1_2-1-1	221	5.85e-20	Q E Q S G T A G S L K Y S D K S D V S F G M V S F E L L T G K V P F E D S H L Q G D K M S R N I
At1g04210.1_protein	999	1.02e-19	E V F R A M H E Q N F Y G L E V D I W S F G C L I F E L L T L Q N F Y F D L S E L Q I H E S L Q N
At1g05100.1_4-4-1_MAPKK18	180	1.76e-19	F M A P E A A R G E R Q G K E S D I W A V G C T V I E M V T G S Q P W I G A D F T D P V S V L Y R
At1g05700.1_1-8-1	737	1.17e-18	Y L D P L C F E T N G L N E K S D I Y S F G V I L E M I T G K T V I K E S Q T K R V H V S D W V
At1g03740.1_4-5-2	390	1.43e-18	R P E L L L G A S H Y G V G V D L W S T G C I L G E L Y A G K F I L P G K T E V E Q L H K I F K
At1g04700.1_2-1-4-1_Raf16	051	1.58e-18	A P E L L N G S S N R V S E K V D V F S P G I V M N E I L T G E E P Y A N L H C G A I I G G I V N
At1g04440.1_3-1-1-1	189	2.67e-17	Y A S V N T H L G I E Q S R R D D L E S L G Y L I M Y F L R G S L F W Q S L R A G T K K Q K Y D K
At1g03930.1_3-1-1-1_ADK1	189	6.15e-17	Y A S V N T H L G V E Q S R R D D L E A L G Y V I M Y F L K G S L F W Q S L K A G T K K Q K Y D R
At1g02970.1_4-3-1	420	1.98e-16	M P Q E I L N E D Y E H L D K V D I F S L G V T V Y E L I K S F L T E S R N Q S L N I K E G K L
At1g01740.1_1-16-1	225	8.67e-16	F T P P E Y L R T G R I T A E S V I Y S F G T L L L D L L T G K H I P P S H A L D L I R D R N L Q

Figura 11.42. Información sobre los sitios

Finalmente, para cada motivo se muestran los diagramas de bloques del motivo en el conjunto de secuencias de entrenamiento. Cada diagrama se corresponde con una secuencia y se muestran todos los sitios que han sido implicados en ese motivo y en esa secuencia, y se muestran también estas en el mismo orden en que se introdujeron en la configuración de la búsqueda.

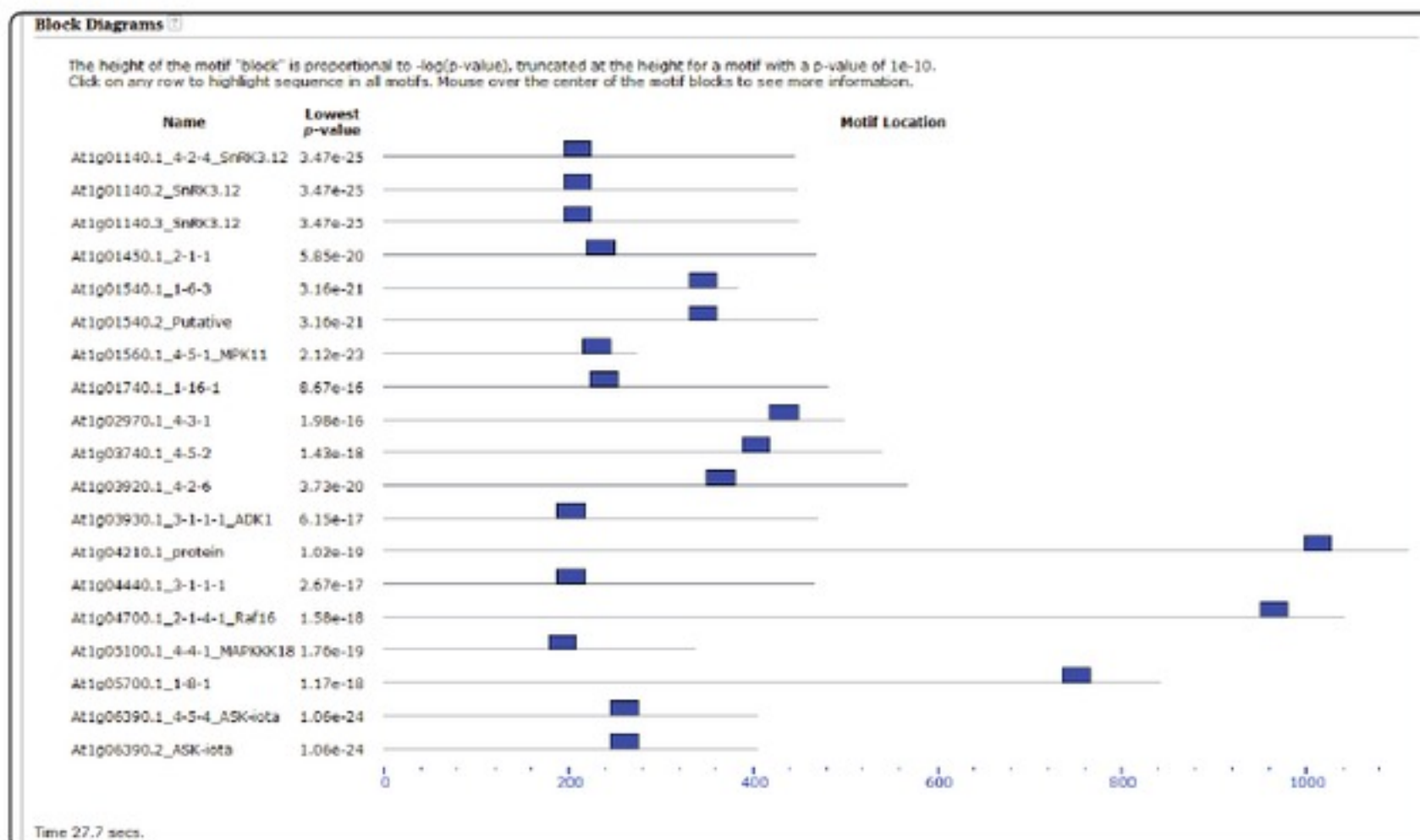


Figura 11.43. Diagramas de bloques de un motivo

Al final de la página de resultados encontramos un diagrama combinado de bloques, en el que se observan las ocurrencias de todos los motivos (ver Figura 11.44).

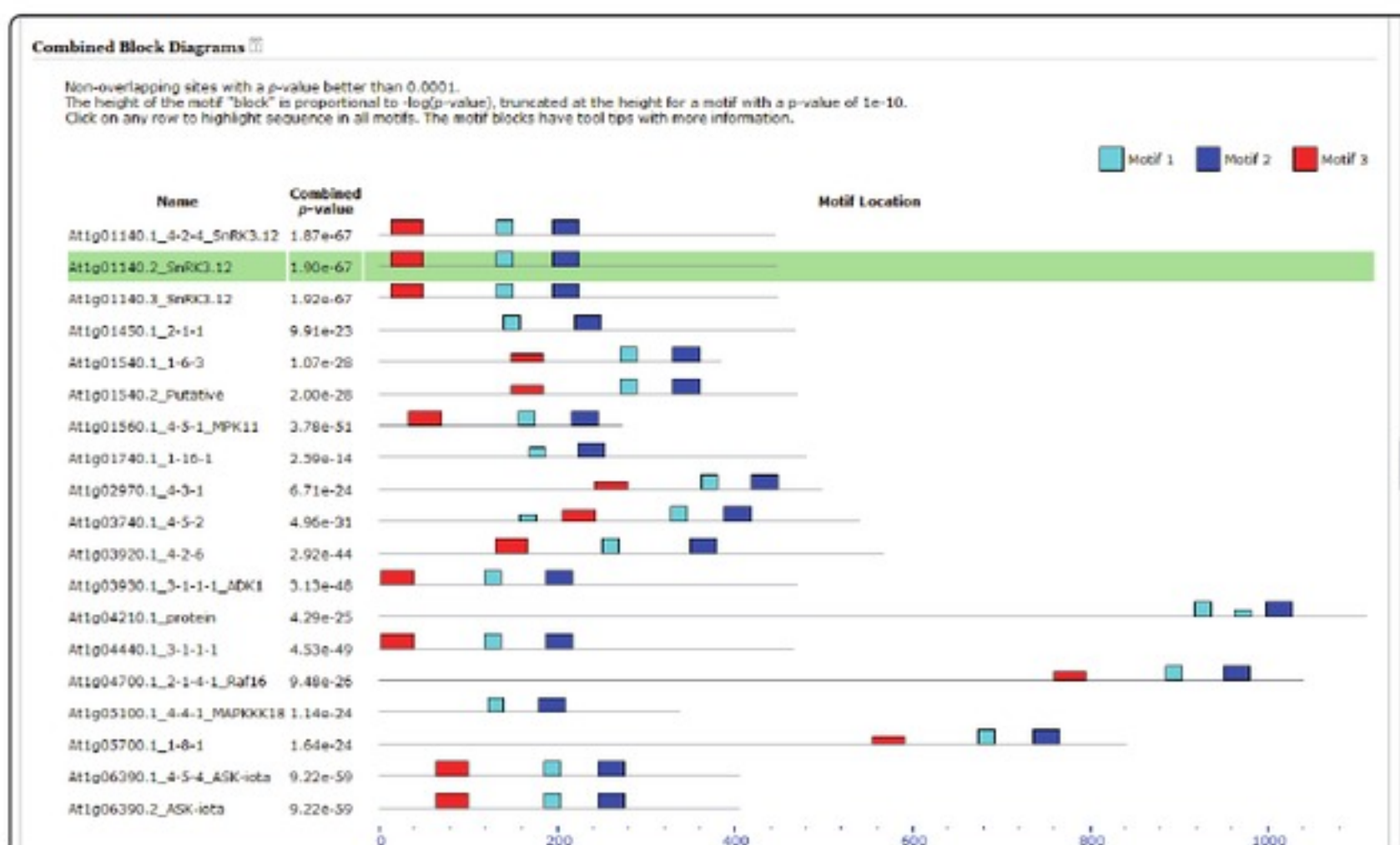


Figura 11.44. Diagrama de bloques combinado



Puede ocurrir que el diagrama de bloques combinado no se corresponda exactamente con la combinación de los diagramas de bloques de cada motivo, puesto que solo se muestran los motivos con un valor p de 0,0001 que no se solapan con otros

Una vez que se han identificado los posibles motivos en nuestras secuencias, debemos comprobar si están presentes en otras secuencias. Esto se consigue gracias a la herramienta de la *suite* MEME llamada MAST, y que está accesible desde la sección **Further Analysis** (ver Figura 11.45):



Figura 11.45. Sección Further Analysis

Esto nos llevará a la página principal de MAST (ver Figura 11.46), en la que deberemos volver a introducir la dirección de correo electrónico en donde queremos recibir los resultados y una descripción de los motivos, junto con información adicional cuyo significado explicamos en los siguientes párrafos.

Figura 11.46. Página principal de MAST

Inicialmente, tenemos dos filtros relacionados con el valor E. El primero de los filtros es el filtro de visibilidad de secuencias (**Display sequences with E-value below**) que elimina de los resultados todas las secuencias cuyo valor E se encuentre por debajo del valor umbral que indicamos en este filtro (ver Figura 11.47).

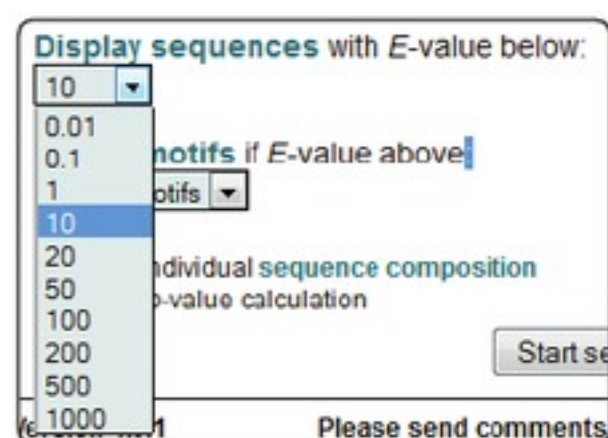


Figura 11.47. Filtro de visibilidad de secuencias

El otro filtro de valor E es el filtro de motivos, que indica si deseamos ignorar los motivos con un valor E mayor que el umbral seleccionado (ver Figura 11.48).



Es recomendable establecer un valor en este filtro, puesto que los motivos con valores E elevados es probable que resulten biológicamente poco significativos.

El umbral por defecto hará que MAST utilice todos los motivos (**use all motifs**), independientemente del valor E de cada motivo.

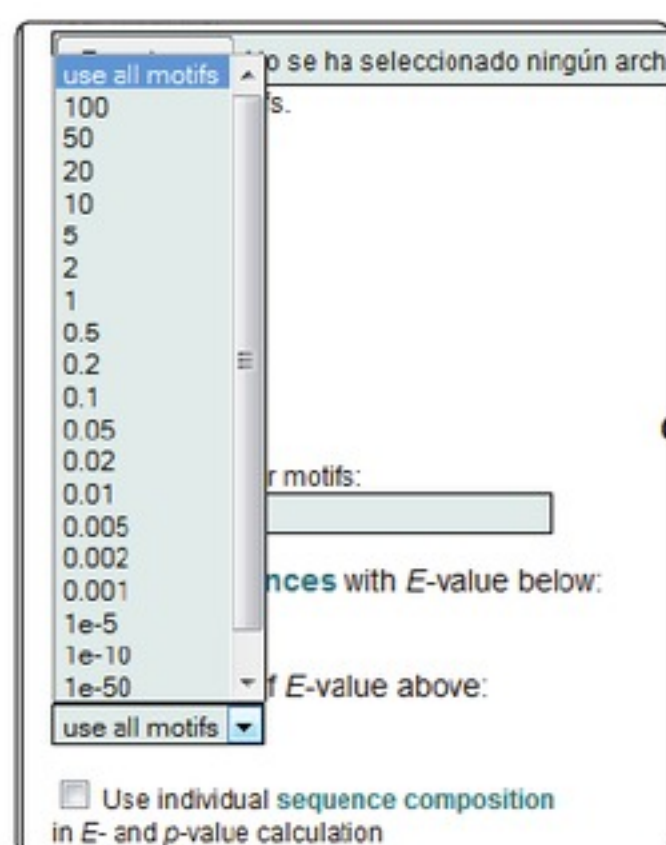


Figura 11.48. Filtro de valor E para los motivos

Si marcamos la casilla **use individual sequence composition in E-and p-value calculation**, mejoraremos la búsqueda selectiva cuando se produzcan coincidencias erróneas debidas a la composición de las secuencias. Generalmente, MAST calcula el valor E y el valor p utilizando un modelo de secuencia aleatorio basado en todas las letras de la base de datos sobre la que se realiza la búsqueda. Con esta opción, se utilizará un modelo aleatorio diferente y basado en la composición de la secuencia objetivo para cada una de estas secuencias, en lugar de la composición de toda la base de datos.



Marcar esta opción puede incrementar, considerablemente, el tiempo de búsqueda debido a la mayor carga de proceso que supone.

Además, debemos indicar la base de datos en la que se buscará el motivo. Para ello, existen dos desplegables (**supported database category** y **database**). En nuestro caso, seleccionaremos **Non-redundant and Special Databases** en la primera lista y **swissprot** en la segunda.

Required

Sequence database to search--select one of the following:

A supported database:

Category:

archivo.

Non-redundant and Special Databases
Other Genomes
ENSEMBL Genomes
ENSEMBL Ab Initio Predicted Proteins
GENBANK Fungi Genomes
GENBANK Bacteria Genomes
Upstream Sequences

Clear

Sample DNA database.

gún archivo.

Figura 11.49. Bases de datos de búsqueda



La lista disponible en http://meme.nbcr.net/meme/cgi-bin/get_db_list.cgi?db_names=fasta_db.csv&mode=doc&short_only=1 describe todas las bases de datos en las que busca MAST.

Es posible escalar el umbral de visualización de motivos por la longitud de la secuencia (**scale motif display threshold by sequence length**). Por defecto, MAST muestra los motivos cuya puntuación se encuentra por encima de un cierto umbral y calcula el umbral basándose en la probabilidad de los motivos, teniendo en cuenta la longitud media de las secuencias de proteínas, sin tener en cuenta la longitud de la secuencia. Seleccionando esta opción, el umbral de visualización se calcula considerando la longitud media de las secuencias, con lo que se reducen el número de motivos mostrados y el tamaño del fichero de salida.

Finalmente, es posible lanzar las búsquedas de la versión, en secuencias de bases, de la secuencia de aminoácidos de la proteína.

Una vez establecidos los parámetros de la búsqueda, pinchando en el botón **Start search**, obtendremos, al igual que en el caso de MEME, las páginas intermedias de resultados hasta que, una vez concluido el procesamiento, se muestra el resultado final (ver Figura 11.50), en donde seleccionaremos la versión html.

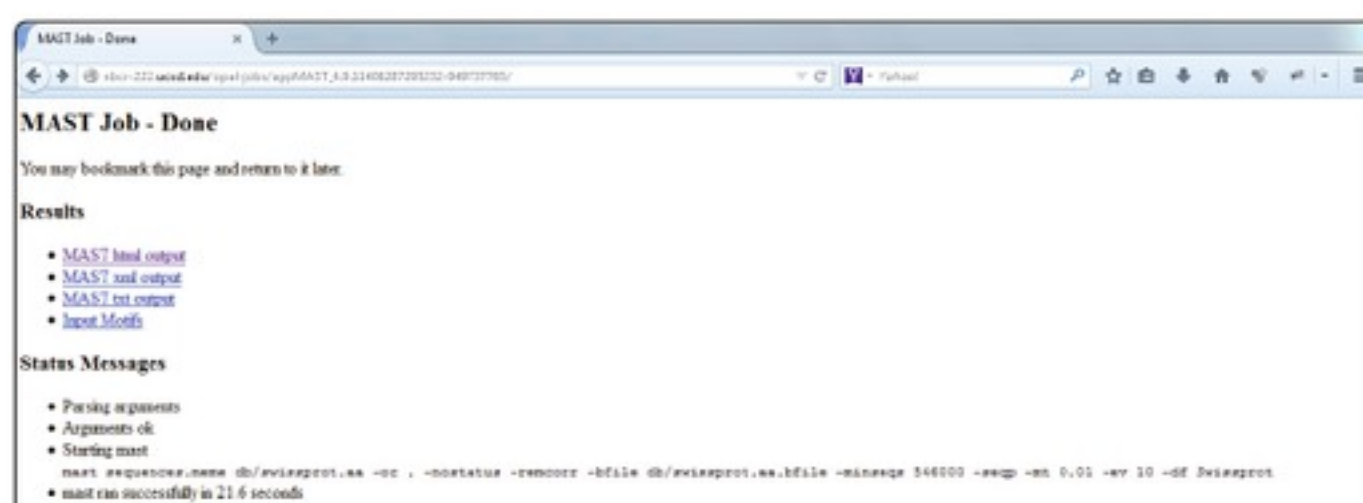


Figura 11.50. Página de resultados

Los resultados de MAST son más sencillos de interpretar que los de MEME. Tras el resumen de la información de entrada a MAST (ver Figura 11.51), encontramos los resultados de la búsqueda.

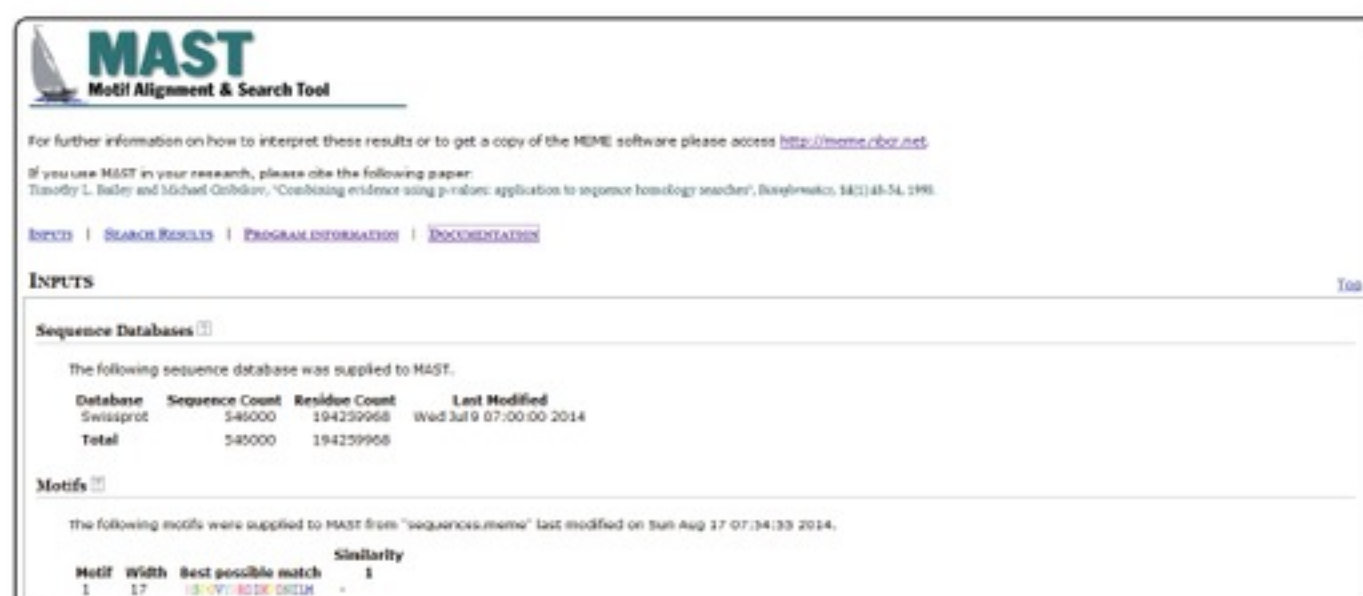


Figura 11.51. Resumen de la información de entrada a MAST

Las secuencias encontradas (ver Figura 11.52) se ordenan por el valor E de la secuencia, de menor a mayor. Tras el nombre de la secuencia y su valor E, aparece una flecha cuya pulsación muestra información adicional de la secuencia en un panel que se abrirá con dicha pulsación. A continuación, podemos ver el diagrama de bloques de los motivos que se han encontrado en las secuencias de la base de datos. En cuanto a la información adicional, se muestra la descripción de la secuencia, el valor p combinado y la secuencia anotada, indicando la posición del motivo dentro de la misma.

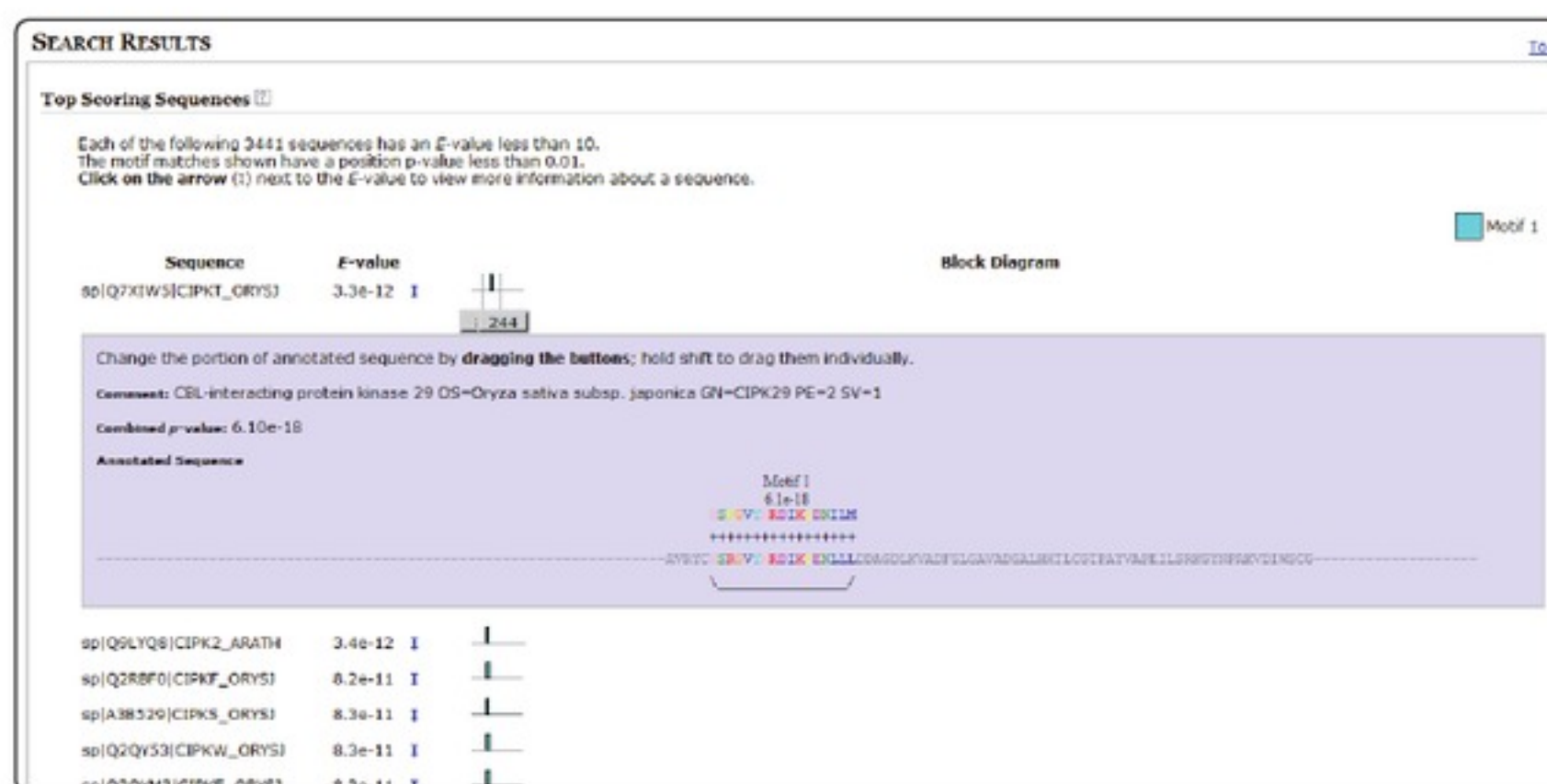


Figura 11.52. Resultados de la búsqueda en MAST

11.6 ANÁLISIS DE LA ESTRUCTURA PRIMARIA DE UNA PROTEÍNA

La estructura primaria de una proteína se considera caracterizada por la secuencia de aminoácidos que la constituyen y el orden en que se disponen para formar la molécula. Conocer la estructura de una proteína permite, no solo conocer su influencia en el comportamiento físico-químico del organismo, sino que sienta las bases para averiguar cómo interferir en ellas e incluso, diseñar proteínas “a medida”.

11.6.1 Traducción del ADN en secuencia proteica

Una vez que se ha obtenido la secuencia de ADN del gen, y una vez garantizada la calidad de dicha secuencia tras haber estudiado la contaminación, el paso siguiente es la traducción del ADN en una secuencia de aminoácidos.

En este primer ejemplo, utilizamos la herramienta de traducción de ExPASy (<http://www.expasy.org/tools/dna.html>) y trabajaremos con la secuencia:

```
>IR64 Oxox2
gcctctacggggtcttgactgctgactaatcttcctataagttaattgcaaattttctca
cgtgtgccacgtacgcatgaaacatgatcagttatgcacggaaggcacgatcgatgcatg
ggtgctcccctatataaagggtccaaagctaactagatcatcagcaaagcaaagtagca
aacaagccagccagctcgtcactgctactgtcttgctgattgaagaagtaattagtta
ctagtagttgattagcaatggagtagcggttcaaagcagctgggttggtgttcgtcgtgc
tgctcctgcagcaggcgcccggtgtaataccgagccaccgacgcggaccctctgcaggatt
tctgcgtcgtgacctcaacagcgaggtgacgggtgaacggggcacgcgtgcaagccggcgt
cgcccgccggcgacgagttcctcttctcctccaagattgccacggggcggcgacgtgaacg
ccaacccgaacgggtccaacgtcacggagctcgacgtcgccgagtgggccggcggtcaaca
cgctcggcggtgtccatgaaccgcgtcgacttcgcgcccgggtggcaccacccggcgacg
tccacccgcgcgcaccgaggtcggtcatcgtgctccgcggcgagctcctcgtcggcatca
tcggcaccctcgacaccgggaacaggtactactccaaggtgggtccgtgccggcgagacgt
tcgtcatcccgaggggggtcatgcacttccagttcaacggtggcaagacggaggccacca
tggtgggtgtccttcaacagccagaaccccggtcgtcttcgtcccgcgtcacattgttcg
gctccaacccgcccattcccgacgcgggtgcttgtcaaggcactccgcgtggatgctgggtg
tagttgagctgctcaagtccaaattcacccggcggtactaattaatctgggagtattttc
gtaccggttcttcgctacgagcagcgtac
```

Figura 11.53. Secuencia bajo estudio

Una vez copiada en el área de texto, pulsamos el botón **Run** y obtendremos la página de resultados (ver Figuras 11.54 y 11.55):

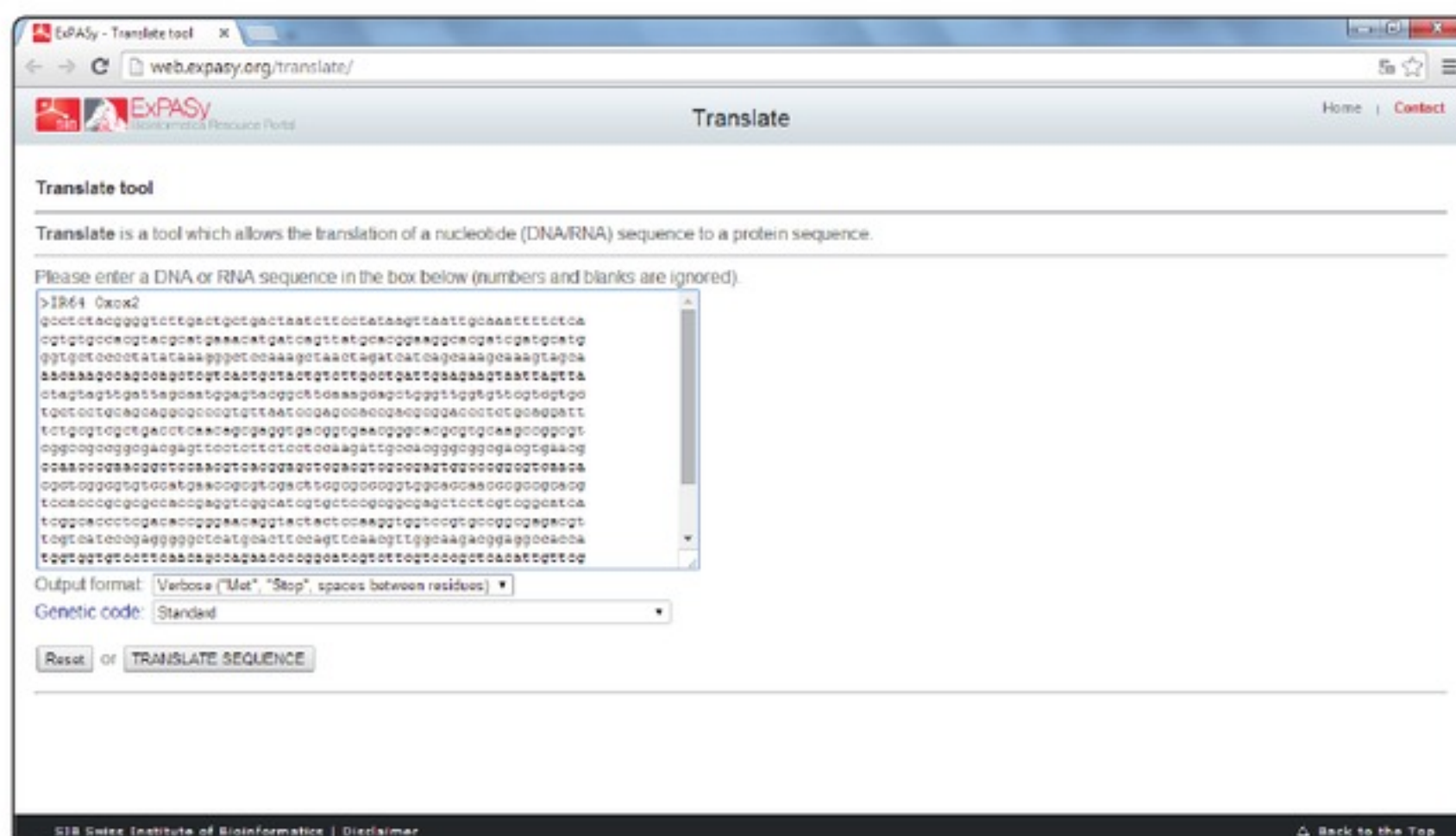


Figura 11.54. Página principal de ExPASy Translate Tool

En este caso, los resultados son 6 tramas (tanto directas como inversas) de la secuencia traducida que muestran Met (metonina), el codón de inicio y el de parada en negrita.

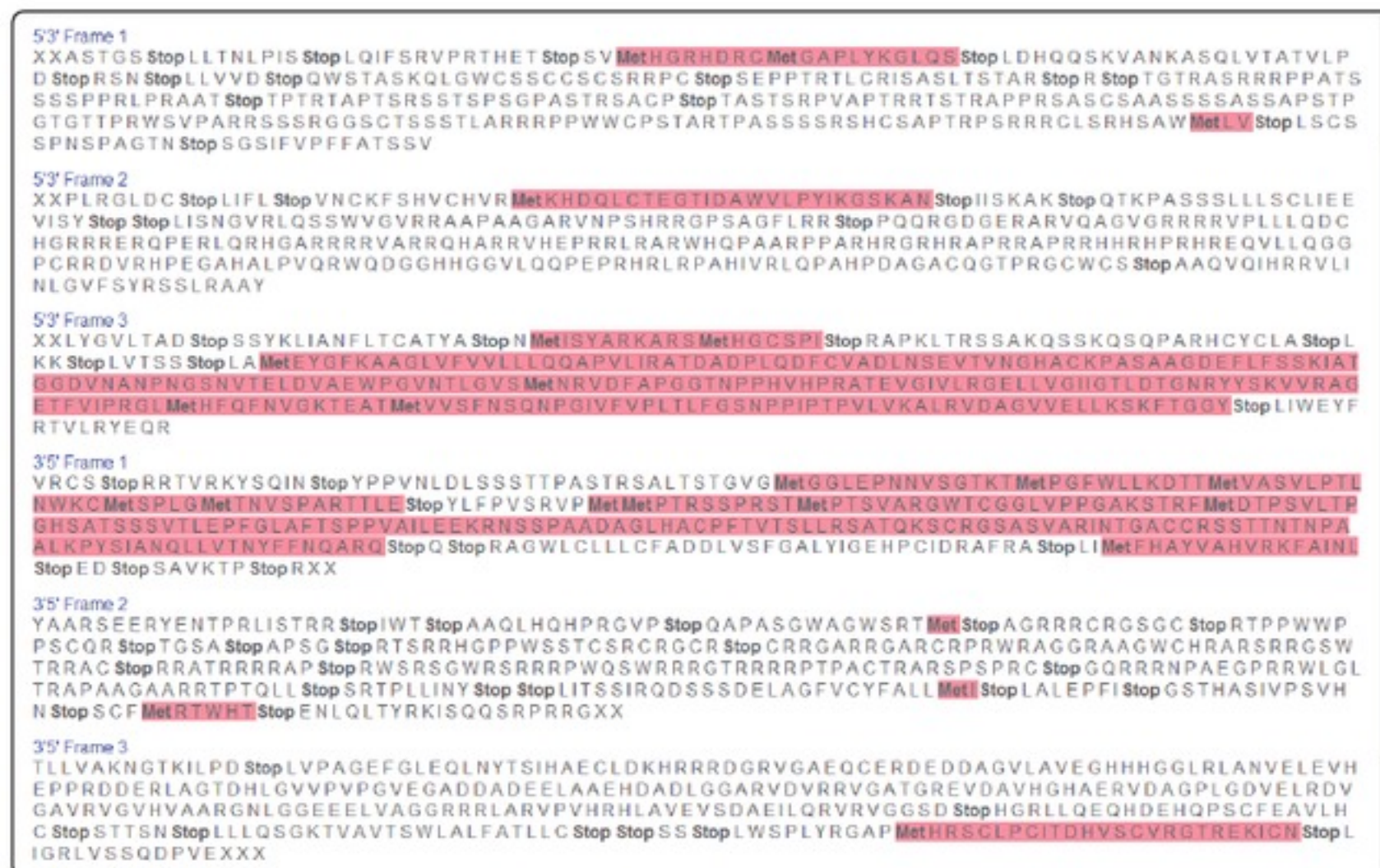


Figura 11.55. Página de resultados de ExPASy Translate Tool

Generalmente, la región mejor traducida es el fragmento más largo. En este caso, se corresponde con la siguiente secuencia escrita en sentido 5'3' de la tercera trama.

```
MEYGFKAAGLVFVLLLLQQAPVLIRATDADPLQDFCVADLNSEVTVNGHACKPASAAGDEF
LFSSKIATGGDVNANPNNGSNVTELDVAEWPGVNTLGVS MNRVDFAPGGTNPPHVHPRATEV
GIVLRGELLVGIIGTLDTGNNRYYSKVVRAGETFVIPRGLMHFQFNVGKTEATMVVSFNSQN
PGIVFVPLTLFGSNPPIPTPVLVKALRVDAGVVELLKSKFTGGY
```

Figura 11.56. Fragmento más largo

11.6.2 Predicción de las propiedades físico-químicas

La estructura primaria de la proteína establece, entre otros aspectos, las propiedades físico-químicas de la misma. Para su estudio, utilizaremos la herramienta ProtParam de ExPASy (<http://expasy.org/tools/protparam.html>).

ProtParam de ExPASy permite calcular varios parámetros físicos y químicos de una secuencia proteica dada, entre los que se encuentran el peso molecular, el pI teórico, la composición de aminoácidos, la composición atómica, el coeficiente de extinción, etc.

En la pantalla principal de ProtParam copiaremos la secuencia de aminoácidos, aunque también es posible introducir el *accession* de la proteína (por ejemplo, P05130) o el identificador de la secuencia (por ejemplo, KPC1_DROME).

```
MCGGEVLPADMPAAPFTPRHGDGETWVDRKRRNKKRKRGADEEWEAAFQEFMAADDDDDGGGLV
LSSKSLVLRSPGENDAGRGAAATMSMPLDPVTEEAEPVAEKP RRRRP RRSY EYHGIRQRPWGRW
SSEIRDPVKG VRLWLGTFTAVEAALAYDAEARRIHGWKARTNFPPADLSSPPPPPQPLCFLND
NGLITIGEAPTDDAASTSTSTTEASGDARIQLECCSDDVMD SLLAGYDVASGDDIWTWTSGASST
SVNQEIKTPSIHQNI SYAGARPMTCHFKNHKNTFVQMECSTMLNLLKGHKQ
```

Figura 11.57. Secuencia de aminoácidos de la proteína P05130 de la mosca de la fruta

En el caso de indicar un *accession*, se muestra una pantalla intermedia previa a la pantalla de resultados (ver Figura 11.58), en donde es posible seleccionar para el análisis o bien la secuencia completa o bien los dominios funcionales que se hayan encontrado.

ProtParam

Selection of endpoints on the sequence

KPC1_DROME (P05130)

Protein kinase C, brain isozyme (EC 2.7.11.13) (PKC) (dPKC53E(BR))
Drosophila melanogaster (Fruit fly)

Please select one of the following features by clicking on a pair of endpoints, and the computation will be carried out for the corresponding sequence fragment. By default, the complete sequence is used.

Note: Only the features corresponding to subsequences of at least 5 residues are highlighted.

FT	CSAIN	1-679	Protein kinase C, brain isozyme.
FT	DOMAIN	101-378	C2.
FT	DOMAIN	350-608	Protein kinase.
FT	DOMAIN	609-679	AGC-kinase C-terminal.
FT	2K_FING	45-104	Phorbol-ester/DAG-type 1.
FT	2K_FING	119-169	Phorbol-ester/DAG-type 2.
FT	KP_BIND	356-364	ATP (By similarity).

Or, if you wish to select a different sequence fragment (at least 5 amino acids long), you can enter the desired endpoints on the sequence here (by default, the computation will be carried out for the complete sequence).

N-terminal:

C-terminal:

The sequence KPC1_DROME consists of 679 amino acids.

Figura 11.58. Pantalla intermedia

Pinchando en **Submit** se muestra la página de resultados (ver Figura 11.59). En la parte superior de la misma, existe un enlace con el texto del identificador de la proteína que lleva a la información sobre la misma almacenada en Uniprot (ver Figura 11.60).



Figura 11.59. Parámetros calculados por ProtParam

Además del peso molecular o la composición atómica, los parámetros más importantes son:

- ▀ **Coeficiente de extinción:** muestra cuánta luz absorbe una proteína a una cierta longitud de onda y resulta muy útil en los estudios de espectrometría. ProtParam ofrece una estimación que, en todo caso, debe confirmarse experimentalmente.
- ▀ **Inestabilidad:** se trata de una estimación de la estabilidad de la proteína. Cuando el valor es de 40, la proteína suele ser estable. Cuanto mayor sea el valor, más inestable será.
- ▀ **Vida media:** es una predicción del tiempo que tarda la proteína completa en desaparecer después de su síntesis en la célula.

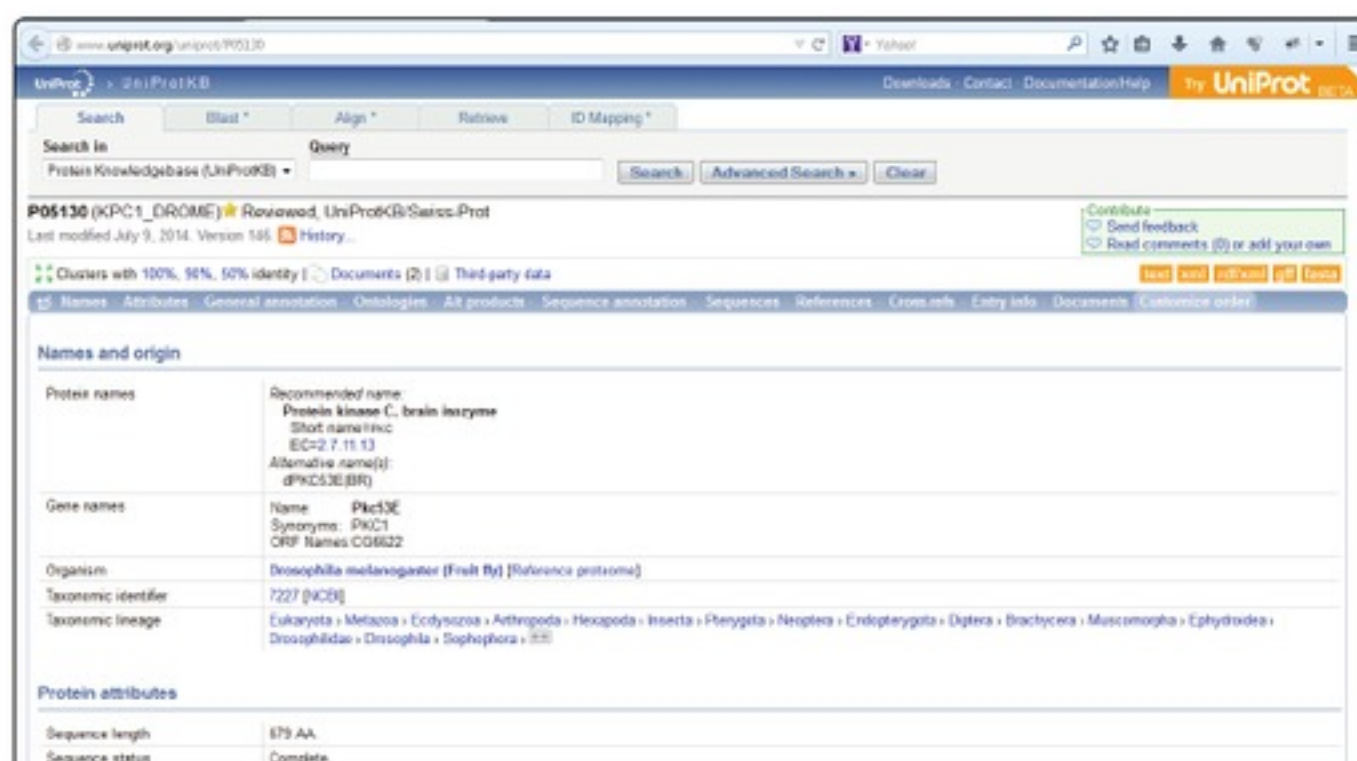


Figura 11.60. Información sobre la proteína en UniProt

11.7 PREDICCIÓN DE LA ESTRUCTURA SECUNDARIA

La estructura secundaria hace referencia al plegamiento de la cadena de aminoácidos en el espacio y está formada por una estructura central y cadenas laterales que se extienden en tres tipos de disposiciones diferentes: por fuera de la hélice (hélice alfa), en hoja plegada (plegamiento beta) o en giros beta (secuencias cortas que imponen un grupo de 180 grados a la cadena principal). Inicialmente, se utilizaron métodos totalmente empíricos que se basaban en el análisis de las estructuras conocidas de algunas proteínas y a partir de los cuales se elaboraron unas tablas que indicaban la frecuencia relativa con la que un cierto residuo de aminoácido se traduce en una hélice alfa, un plegamiento beta o un giro beta (en la Bibliografía pueden encontrarse).



Si la proteína bajo estudio tiene un número suficientemente elevado de homólogos, la predicción de la estructura secundaria tendrá una fiabilidad estimada alrededor del 80%. Sin embargo, no hay que perder de vista que se trata de una predicción que habrá que confirmar experimentalmente.

Una de las herramientas más exactas para predecir la estructura secundaria de proteínas es PSIPRED (<http://bioinf.cs.ucl.ac.uk/psipred/>).

La primera sección de configuración de la búsqueda es el programa que se desea ejecutar. Se dispone de una gran variedad de herramientas de las que hay que elegir obligatoriamente una, siendo PSIPRED la seleccionada por defecto.

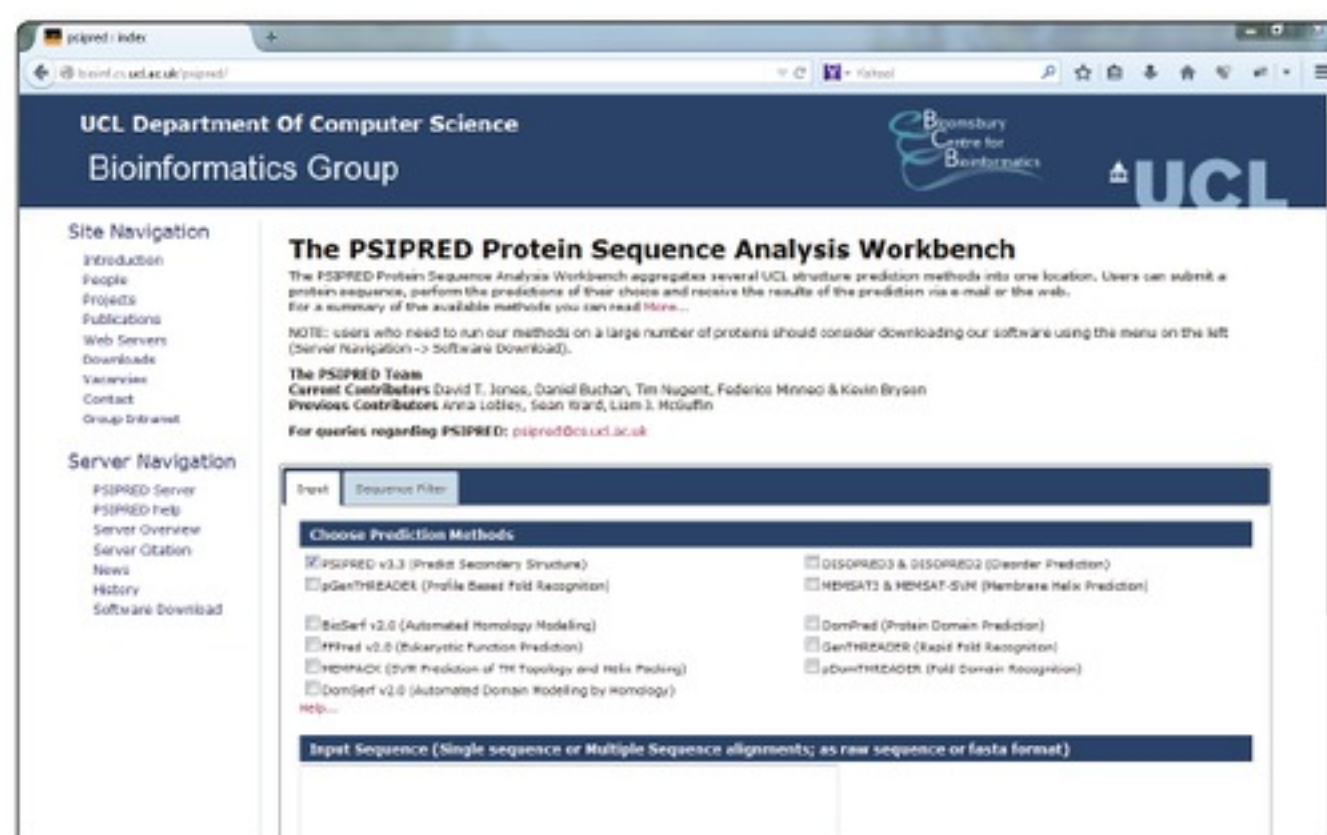


Figura 11.61. Página principal de PSI-PRED



En esta práctica trabajaremos con PSIPRED pero recomendamos que se repita este análisis con otras herramientas distintas y se comparen los resultados.

El paso siguiente es especificar la secuencia bajo estudio. Trabajaremos con la siguiente secuencia proteica (ver Figura 11.62):

```
MEYGFKAAAGLVFVLLQLQAPVLIRATDADPLQDFCVADLNSEVTVNGHACKPASAAGDEFLEFSS
KIATGGDVNANPNNGSNVTELDVAEWPQVNTLGVSMMNRVDFAPGGTNPPHVHPRATEVGIVLRGEL
LVGIIGTLDTGNYYSKVVRAGETFVIRGLMHFQFNVGKTEATMVVSFNSQNPQIVFVPLTLFG
SNPPIPTPVLVKALRVDAGVVLLKSKFTGGY
```

Figura 11.62. Secuencia bajo estudio

Una vez introducida la secuencia en el área de texto **Input Sequence** y, opcionalmente, completados los datos de correo electrónico para recibir los resultados en el buzón de entrada, estamos en condiciones de lanzar la predicción pulsando el botón **Predict**.

The screenshot shows the PSIPRED web interface with the following sections:

- Input** (selected tab) and **Sequence Filter** (disabled tab).
- Choose Prediction Methods** section with checkboxes for:
 - ☒ PSIPRED v3.3 (Predict Secondary Structure)
 - ☐ pGenTHREADER (Profile Based Fold Recognition)
 - ☐ BioSerp v2.0 (Automated Homology Modelling)
 - ☐ FFPred v2.0 (Eukaryotic Function Prediction)
 - ☐ MEMPACK (SVM Prediction of TM Topology and Helix Pecking)
 - ☐ DomSerp v2.0 (Automated Domain Modelling by Homology)
 - ☐ DISOPRED3 & DISOPRED2 (Disorder Prediction)
 - ☐ MEMSAT3 & MEMSAT-SVM (Membrane Helix Prediction)
 - ☐ DomPred (Protein Domain Prediction)
 - ☐ GenTHREADER (Rapid Fold Recognition)
 - ☐ pDomTHREADER (Fold Domain Recognition)
- Input Sequence (Single sequence or Multiple Sequence alignments; as raw sequence or fasta format)** section with a text area containing the protein sequence from Figure 11.62.
- Submission Details** section with:
 - Email Address for job completion alert (optional):
 - Password (only required for licensed commercial e-mail addresses):
 - Short identifier for submission:
- Buttons: **Predict** and **Clear form**.

Figura 11.63. Configuración de los parámetros de búsqueda

Aparecerá una pantalla intermedia (ver Figura 11.64) mientras se está realizando el procesamiento de la información.



Figura 11.64. Procesado del envío

Una vez ha finalizado el análisis, se muestra la página de resultados (ver Figura 11.65), en la que se distinguen tres pestañas.

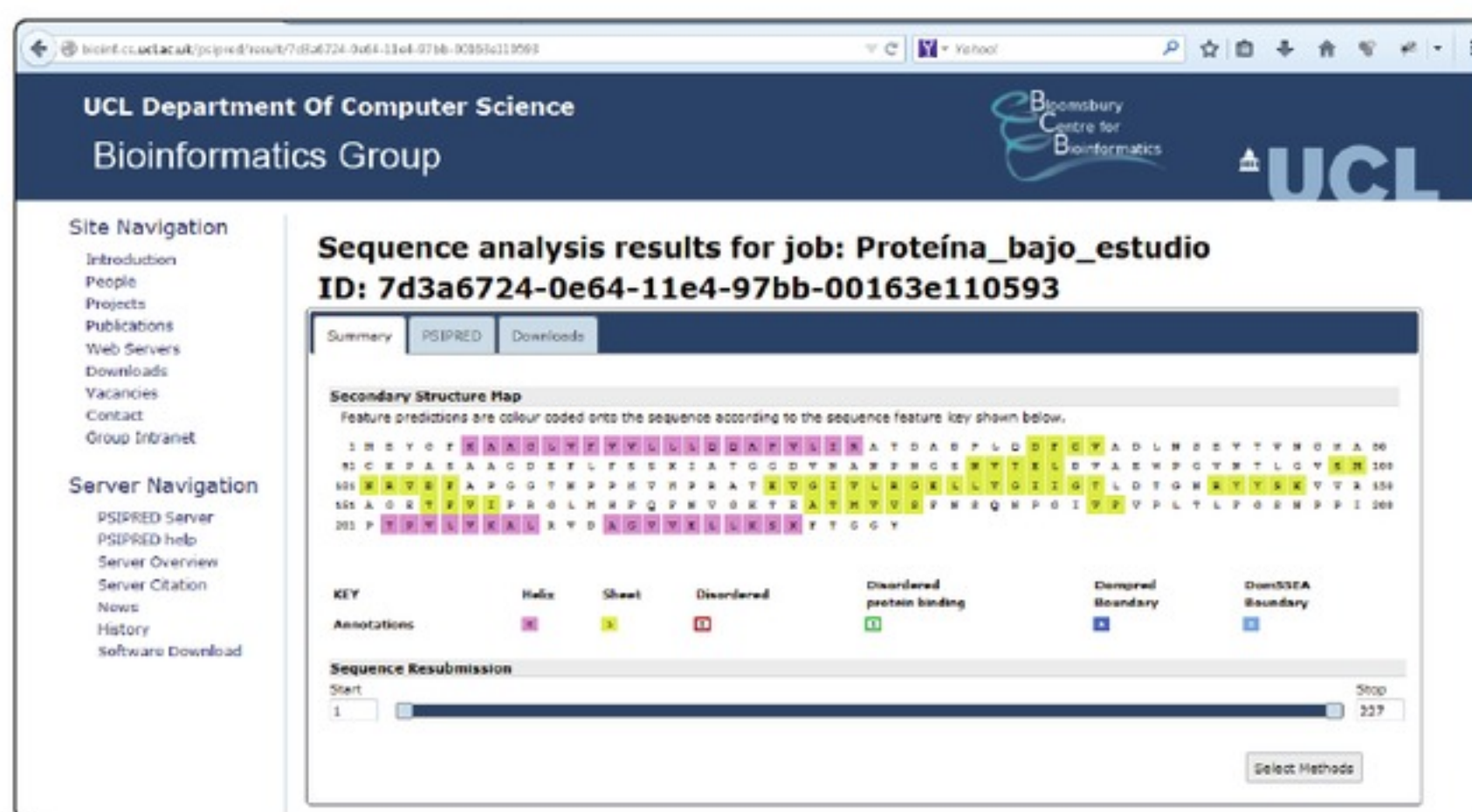


Figura 11.65. Página principal de resultados

En la primera pestaña, encontramos el resumen del análisis de la secuencia en donde se incluyen datos relativos al envío del trabajo al servidor. Sin embargo, la parte más interesante es la correspondiente a la estructura secundaria (ver Figura 11.66).

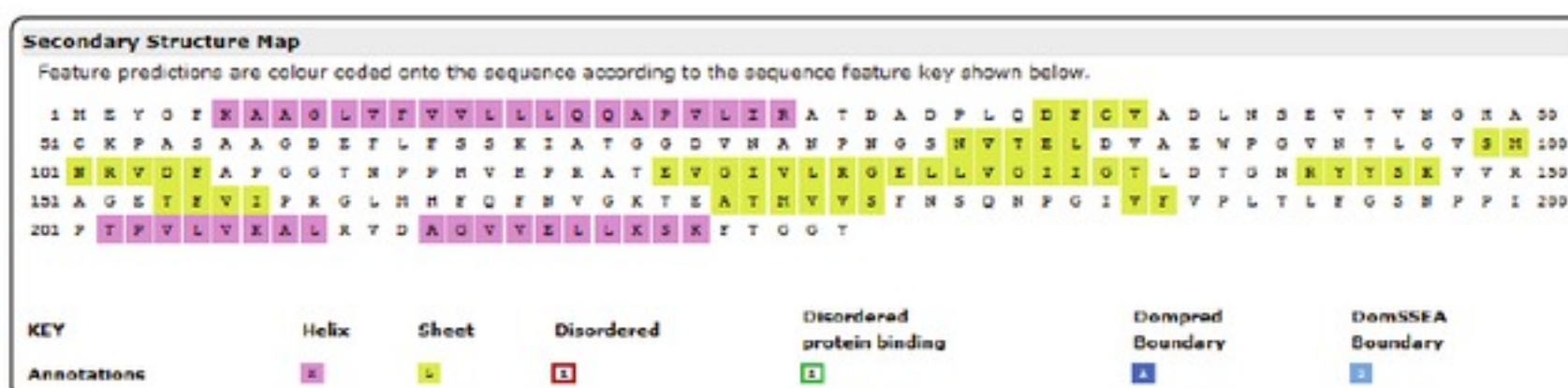


Figura 11.66. Mapa de la estructura secundaria de la proteína

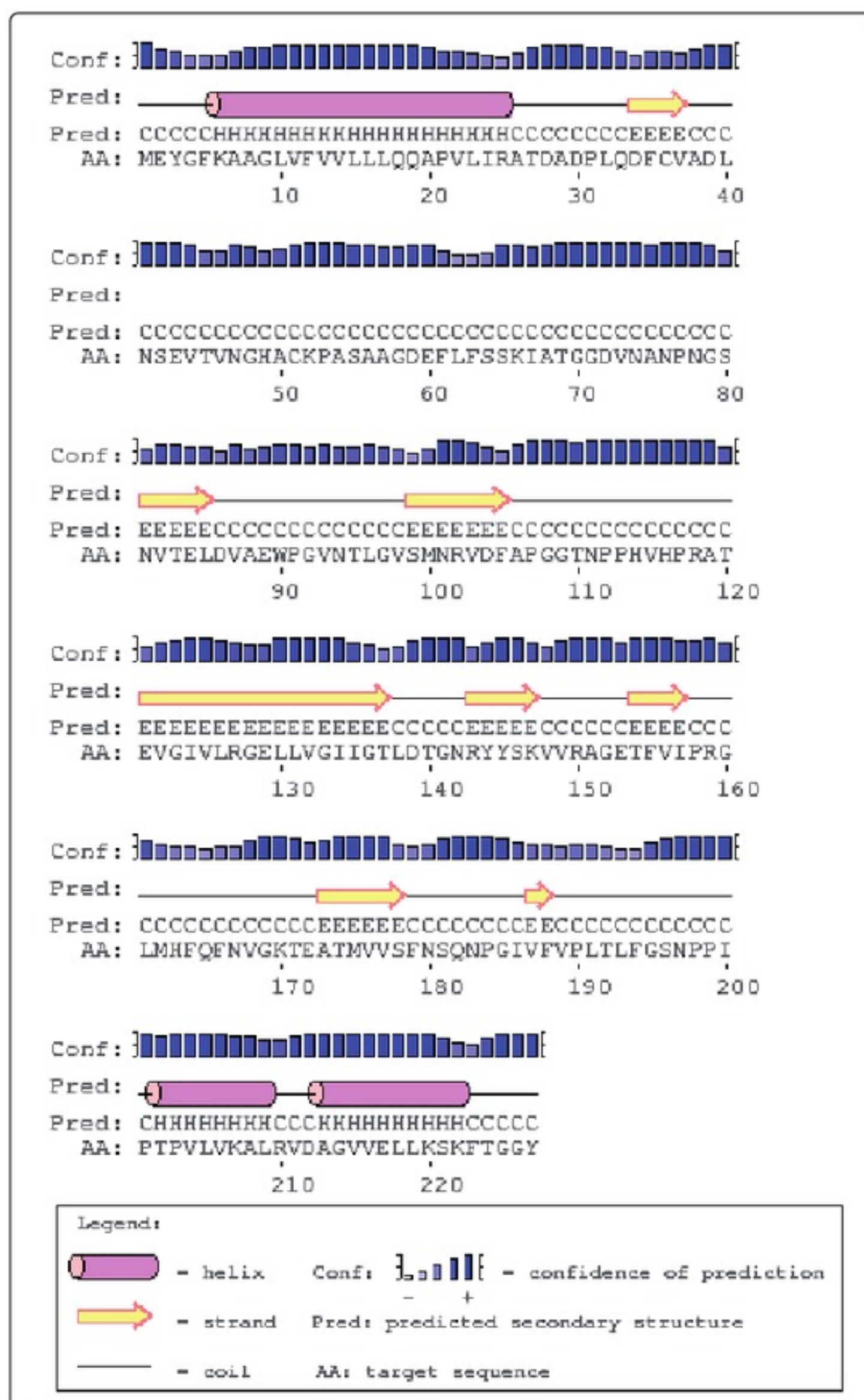
El mapa de estructura secundaria consiste en la secuencia de aminoácidos introducida y formateada en disposición tabular, con un aminoácido en cada elemento siguiendo un código de colores que indica cuál es la posición que adopta en el espacio.

En la segunda pestaña, PSIPRED representa el mapa anterior de una manera más gráfica, aportando información sobre la fiabilidad de la predicción realizada (ver Figura 11.67):



Figura 11.67. Gráfico de resultados

Pinchando en la imagen reducida se muestra la imagen completa (ver Figura 11.68). La línea de predicción (**Pred**) consiste en una sucesión de H (hélices), E (plegamiento beta) o C (giro beta) que indican la disposición espacial del residuo. Por otra parte, la línea de fiabilidad (**Conf**) indica el nivel de fiabilidad de la predicción para cada posición.



Finalmente, es posible descargarse el informe completo en distintos formatos a través de los enlaces que aparecen en la tercera pestaña (ver Figura 11.69):

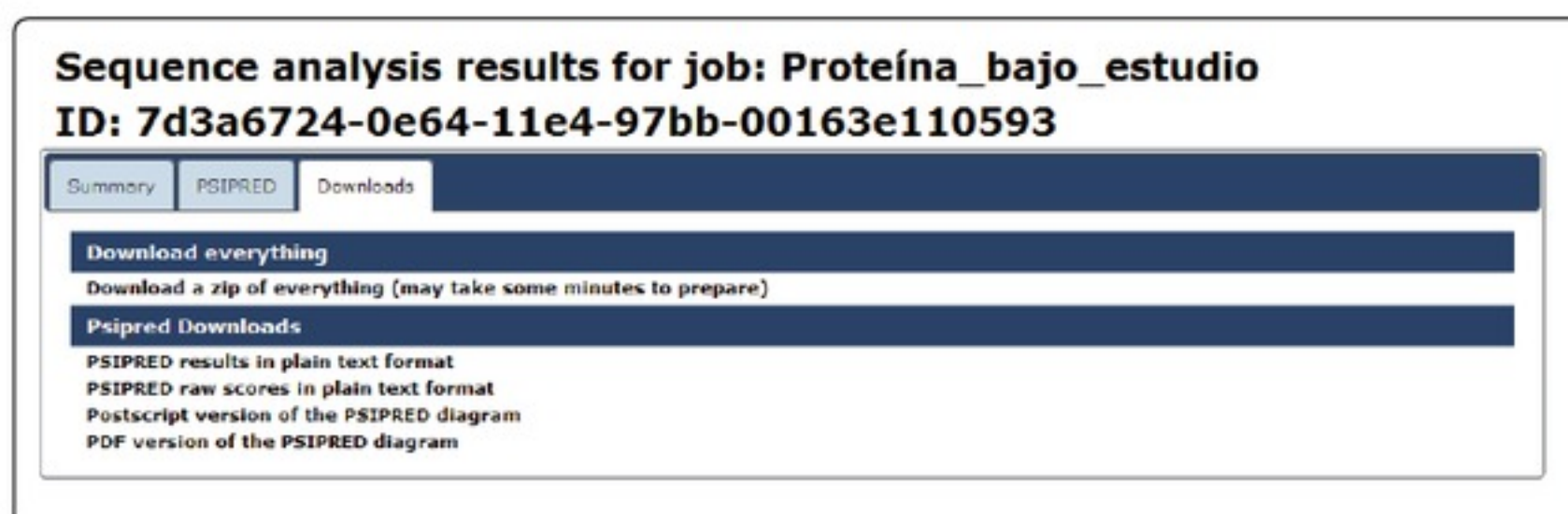


Figura 11.69. Resumen de los resultados

11.8 PREDICCIÓN DE LA ESTRUCTURA TERCIARIA

La estructura terciaria es la estructura tridimensional de la cadena de aminoácidos.

La herramienta Swiss Model es un software de modelado automático de estructuras homólogas de proteínas accesible en <http://swissmodel.expasy.org/>. En la página principal de la herramienta (ver Figura 11.70), pulsamos el botón **Start Modelling**:

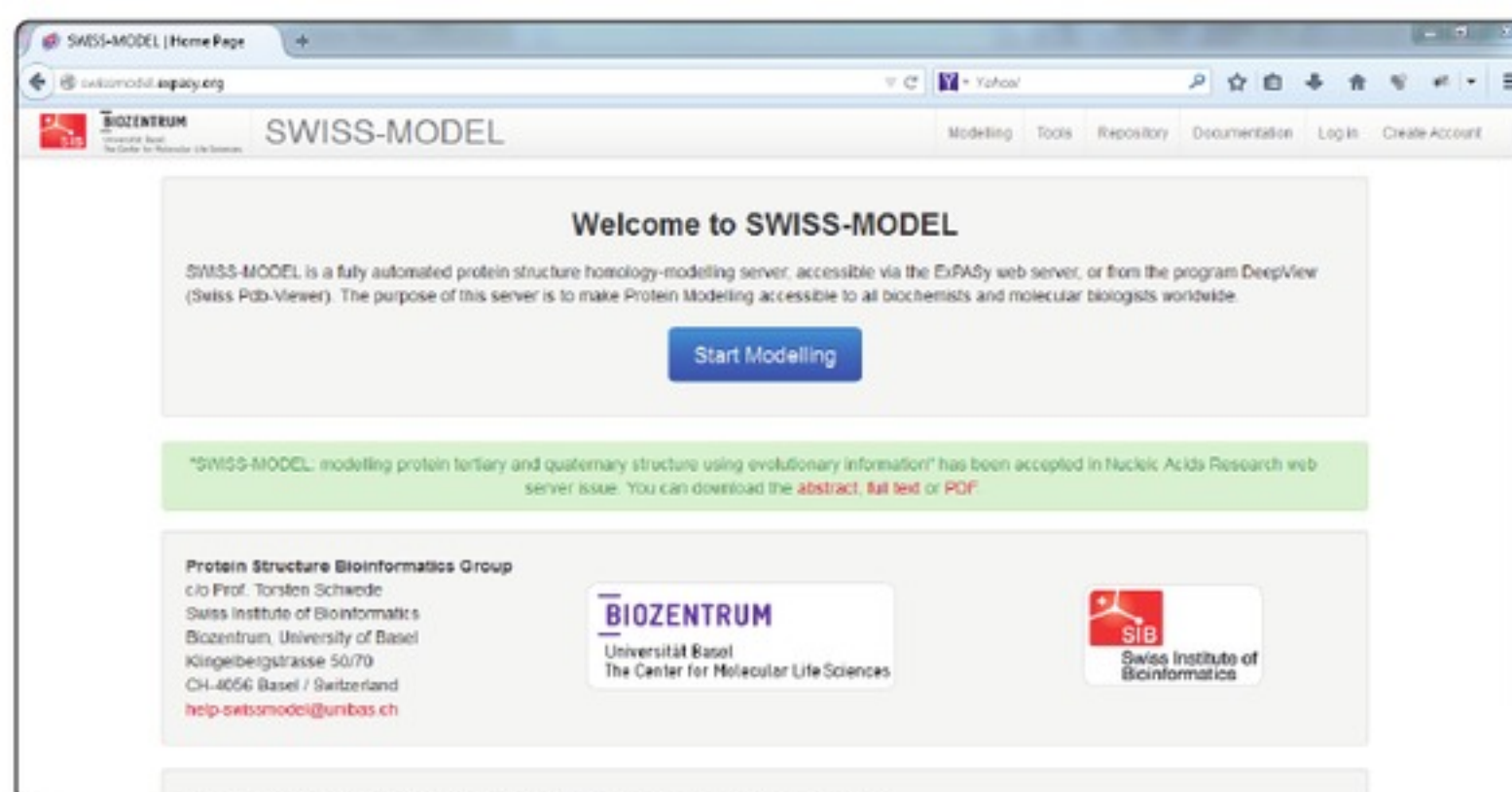


Figura 11.70. Página principal de Swiss-Model

Aparece entonces una pantalla en la que es posible introducir la secuencia correspondiente a la proteína bajo estudio. Trabajaremos con la siguiente (ver Figura 11.71):

```
MEYGFKAAGLVFVLLQLQAPVLIRATDADPLQDFCVADLNSEVTVNGHACKPASAAGDEFLFSS
KIATGGDVNANPNNGSNVTELDVAEWPGVNTLGVSMMNRVDFAPGGTNPPHVHPRATEVGIVLRGEL
LVGIIGTLDTGNYYSKVVRAGETFVIPRGLMHFQFNVGKTEATMVVSFNSQNPFGIVFVPLTLFG
SNPPIPTPVLVKALRVDAGVVELLKSKFTGGY
```

Figura 11.71. Proteína bajo estudio

En cuanto se pega la secuencia, la herramienta la valida y ya se está en condiciones para construir el modelo pinchando en el botón **Build Model**.

Start a New Modelling Project

Target Sequence: (Format must be Fasta, Clustal, Promod, plain string, or a valid UniProtKB AC)

Target: 80

Target: 160

Target: 240

Target: 320

[Reset Form](#)

[Upload Target Sequence File...](#)

Project Title: Untitled Project

Email: Optional

[Search For Templates](#) [Build Model](#)

Figura 11.72. Validación de la secuencia

Los resultados (ver Figura 11.73) muestran la predicción de la estructura tridimensional de la proteína bajo estudio. Aparecen una lista de coincidencias entre la proteína bajo estudio y las referencias guardadas en la base de datos Swiss-Prot. Para cada coincidencia, se proporciona el modelo empleado en el cálculo, el porcentaje de similitud con la proteína bajo estudio, los ligandos y una representación 3D de la estructura de la proteína que puede girarse en el espacio utilizando el ratón.

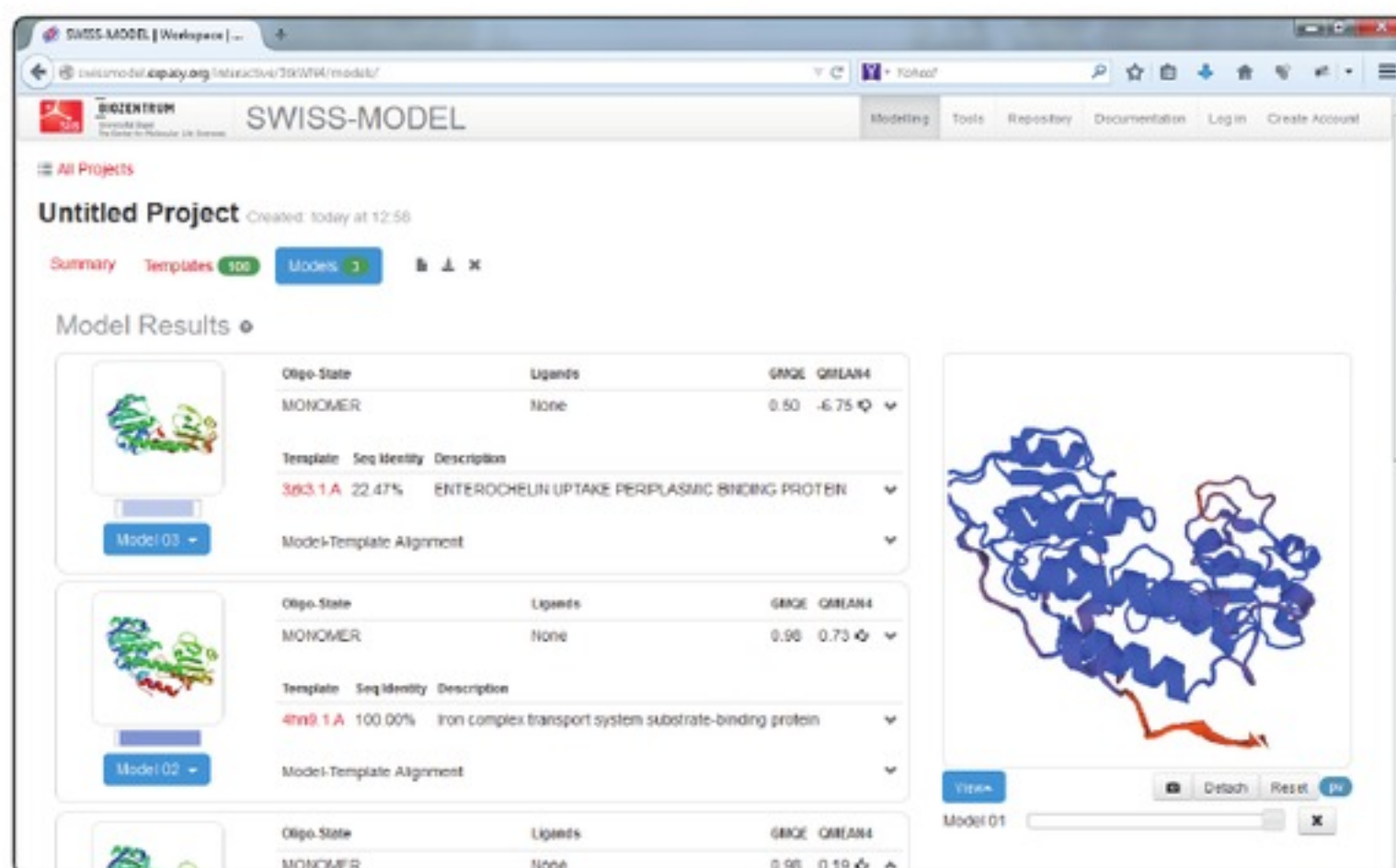


Figura 11.73. Página de resultados

Recientemente, se ha desarrollado un intento de abordar el problema fraccionando el proceso en pasos razonables. El ejemplo más representativo es el programa LINUS (*Local Independently Nucleated Units of Structure*). Se trata de encontrar las regiones con mayor probabilidad de presentar una estructura secundaria determinada y, a partir de ahí, averiguar cómo pueden plegarse juntas para alcanzar una configuración de mínima energía.

11.9 PREDICCIÓN DE GENES CON GENSCAN

Como ya se ha visto a lo largo del libro, los avances tecnológicos han incrementado dramáticamente la velocidad a la que se generan secuencias de ADN. Sin embargo, la caracterización de dichas secuencias continúa siendo un proceso lento e intensivo en recursos. Afortunadamente existen herramientas que facilitan el análisis.

En esta práctica utilizaremos GENSCAN para predecir la localización de genes en secuencias de ADN, si bien no hay que perder de vista que los resultados arrojados por GENSCAN deben ser confirmados experimentalmente.

Siguiendo los ejemplos de las prácticas anteriores, encontraremos la secuencia correspondiente al ADN del cloroplasto del maíz, cuyo *accession* es NC_001666.2. Para ello, abra la página del NCBI (sección **Nucleotides**) y descargamos la secuencia en formato FASTA (ver Figura 11.74):

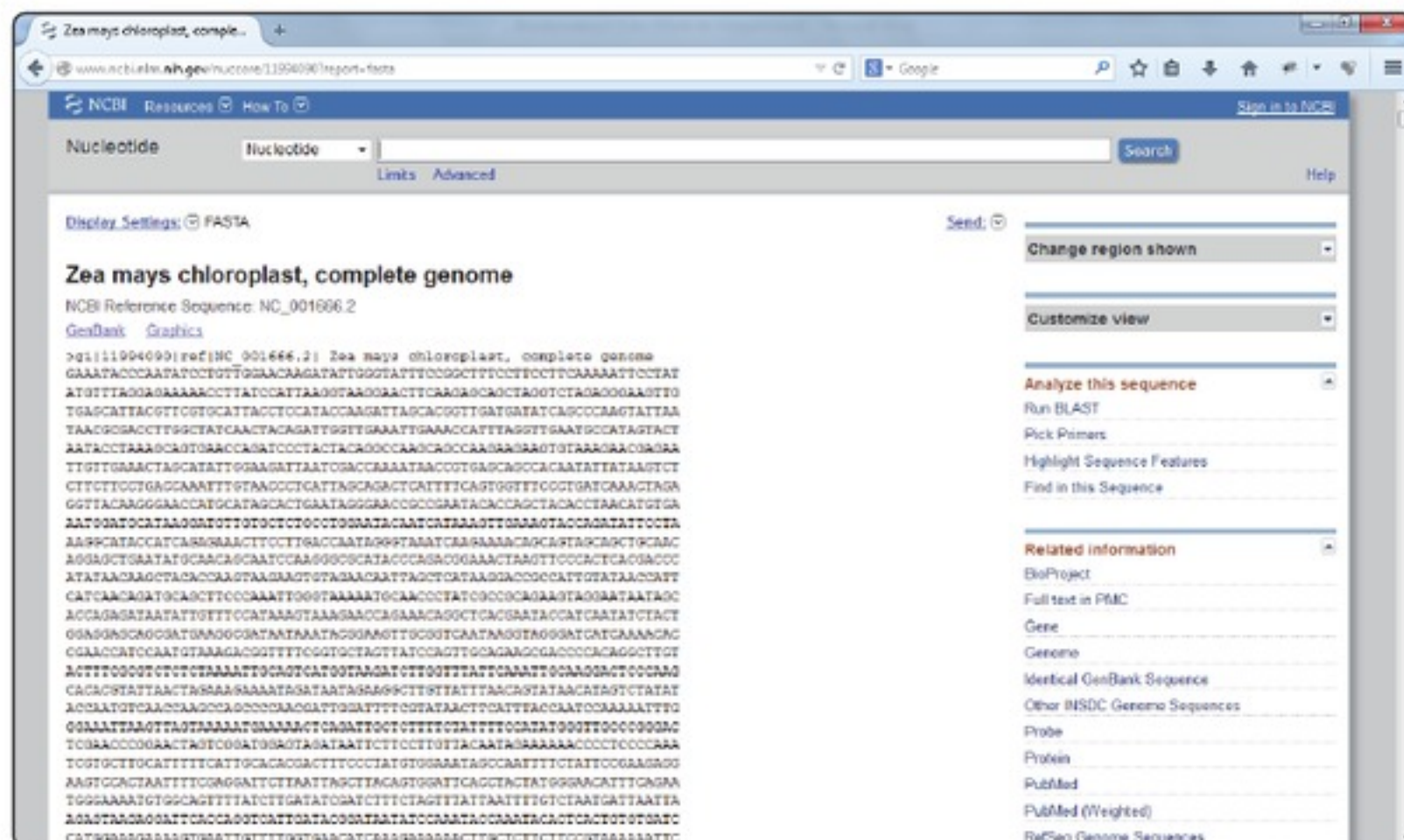


Figura 11.74. Obtención de la secuencia que se desea analizar

A continuación, abrimos la página principal de GENSCAN (<http://genes.mit.edu/GENSCAN.html>) y pegamos la secuencia obtenida en el área de texto acondicionada para ello. También puede descargarse la secuencia en el paso anterior y adjuntar el fichero a GENSCAN.

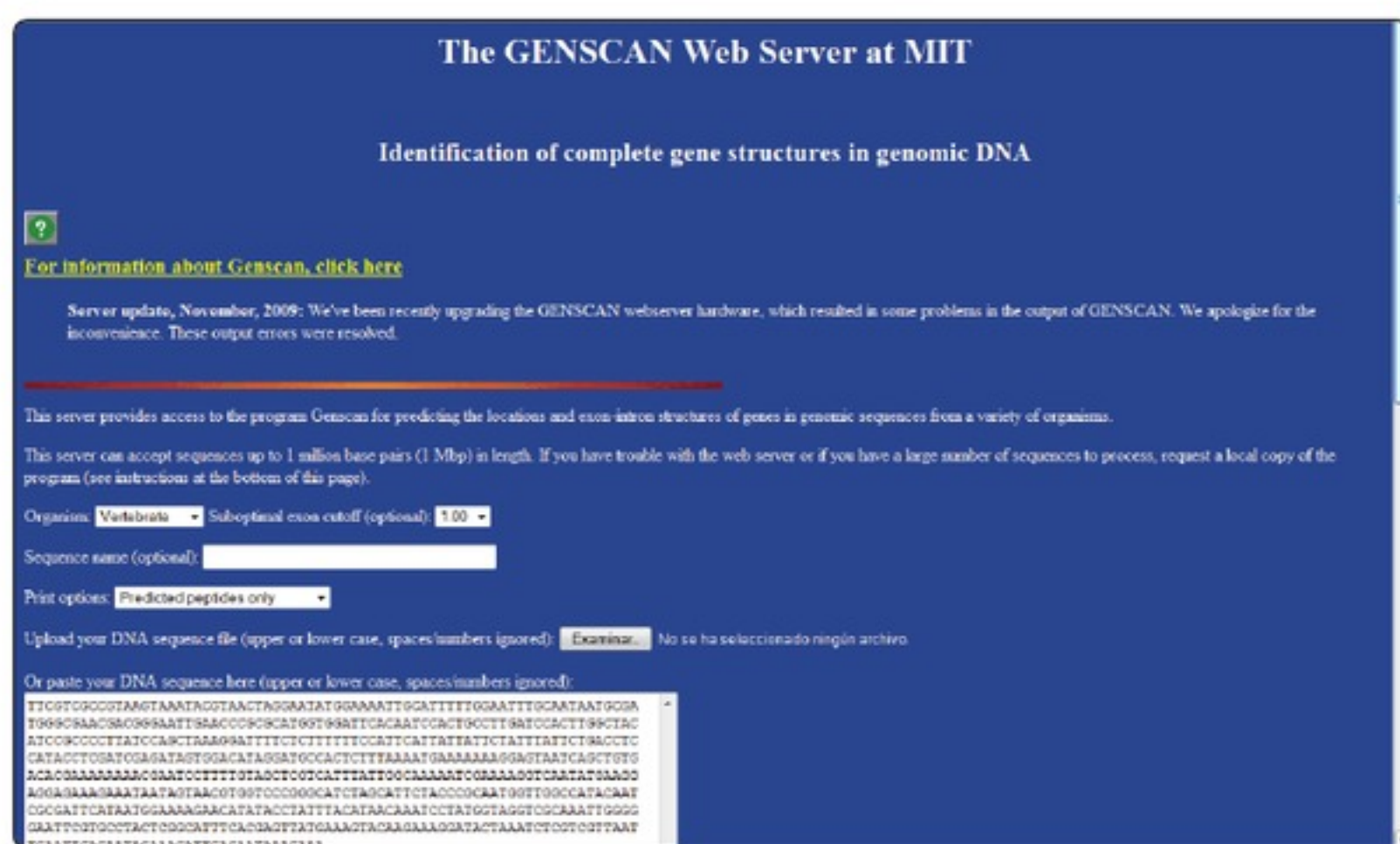


Figura 11.75. Página principal de GENSCAN

En el campo organismo (**organism**) seleccionamos **Maize**. Por otra parte, en las opciones de impresión indicamos **Predicted Peptides Only** (visualiza solamente las secuencias predecidas que codifican péptidos) en lugar de **Predicted CDS and Peptides** (muestra tanto los péptidos como las secuencias codificantes) y pulsamos el botón **Run GENSCAN**.

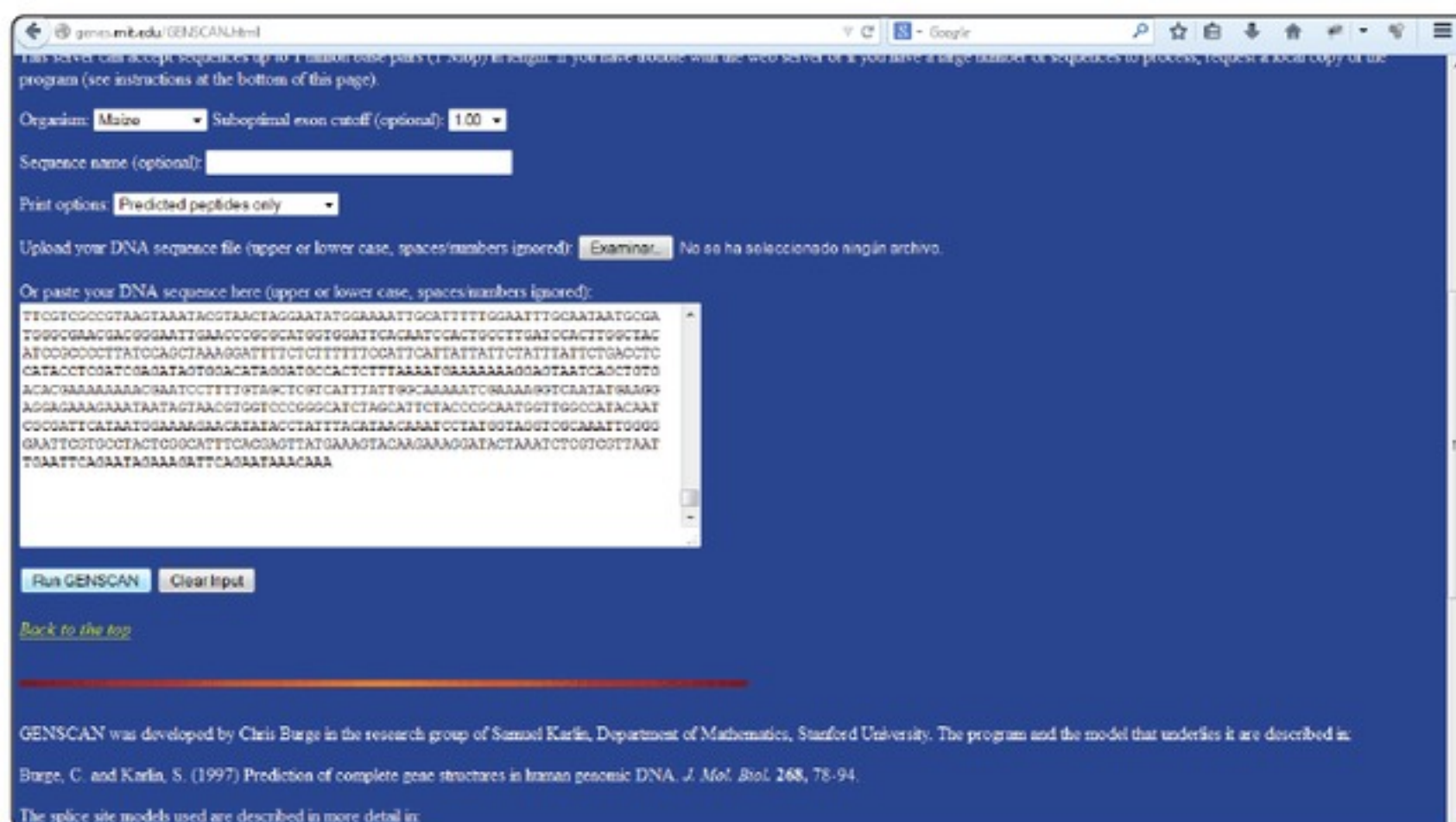


Figura 11.76. Pegado de la secuencia bajo estudio

El resultado es una tabla con información muy detallada de cada gen. En la secuencia bajo estudio existen 12 genes, como se deduce de la columna **Gn**. Es posible averiguar el número de exones de cada gen estudiando la columna **Ex**. Por ejemplo, el primer gen tiene 3 exones.

Las columnas más importantes de la tabla de resultados, sin embargo, son **Type** y **P**. La columna **Type** dice si la predicción es un exón inicial (**Init**), interno (**Intr**), terminal (**Term**), un gen de un único exón (**Sngl**), una región promotora (**Prom**) o una señal PolyA (**PlyA**). Por su parte, la columna **P** indica la probabilidad de que la predicción sea correcta.

```
View gene model output: PS | PDF
GENSCAN 1.0      Date run: 3-Jun-114 Time: 13:05:06

Sequence /tmp/06_03_14-13:05:04.fasta : 140423 bp : 38.46% C+G :
Isochore 1 ( 0 - 100 C+G%)
Parameter matrix: Maize.smat
```

Predicted genes/exons:

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..

1.01	Init	+	1032	1070	39	2	0	110	97	-20	0.601	3.83
1.02	Term	+	7389	7412	24	2	0	84	35	21	0.091	-2.96
1.03	PlyA	+	7562	7567	6						2.27	
2.00	Prom	+	13956	13995	40						3.62	
2.01	Init	+	14148	14153	6	2	0	68	52	10	0.727	-2.01
2.02	Term	+	14443	14448	6	0	0	112	45	0	0.690	-1.72
2.03	PlyA	+	14971	14976	6						2.27	
3.03	PlyA	-	15641	15636	6						2.27	
3.02	Term	-	17123	17111	13	1	1	79	45	18	0.275	-3.37
3.01	Init	-	20228	20218	11	2	2	126	79	21	0.942	6.41
3.00	Prom	-	21496	21457	40						2.72	
4.03	PlyA	-	22600	22595	6						2.27	
4.02	Term	-	22893	22883	11	1	2	107	35	1	0.175	-2.60
4.01	Init	-	28158	28152	7	0	1	86	87	7	0.941	3.36
4.00	Prom	-	31602	31563	40						4.52	
5.00	Prom	+	33254	33293	40						3.82	
5.01	Init	+	34420	34558	139	0	1	27	93	90	0.854	6.30
5.02	Term	+	38378	38385	8	0	2	95	35	0	0.098	-4.25
5.03	PlyA	+	39066	39071	6						2.27	
6.00	Prom	+	39204	39243	40						3.42	
6.01	Init	+	43422	43481	60	2	0	81	52	10	0.349	-1.04
6.02	Term	+	44628	44642	15	2	0	101	46	17	0.266	-0.60
6.03	PlyA	+	45483	45488	6						2.27	
7.00	Prom	+	51967	52006	40						2.82	
7.01	Init	+	52961	52976	16	1	1	48	110	-11	0.255	-0.01
7.02	Term	+	61179	61189	11	1	2	89	45	7	0.046	-2.80
7.03	PlyA	+	61950	61955	6						2.27	
8.00	Prom	+	67861	67900	40						3.32	
8.01	Init	+	71022	71036	15	2	0	35	85	11	0.187	-1.89
8.02	Term	+	81950	81979	30	1	0	79	45	86	0.833	3.71
8.03	PlyA	+	82024	82029	6						2.27	
9.00	Prom	+	85902	85941	40						2.22	
9.01	Init	+	96350	96417	68	1	2	57	84	58	0.053	3.70
9.02	Intr	+	98503	98533	31	1	1	91	72	41	0.140	3.26
9.03	Intr	+	101029	101141	113	0	2	4	114	130	0.821	9.20
9.04	Term	+	102018	102087	70	0	1	-2	46	98	0.274	-3.29
9.05	PlyA	+	104506	104511	6						2.27	
10.00	Prom	+	106533	106572	40						1.42	
10.01	Init	+	111219	111228	10	2	1	68	106	-5	0.549	2.33
10.02	Term	+	116631	116668	38	1	2	71	35	27	0.294	-4.30
10.03	PlyA	+	117050	117055	6						2.27	
11.05	PlyA	-	117698	117693	6						2.27	


```

11.04 Term - 120797 120728 70 1 1 -2 46 98 0.287 -3.29
11.03 Intr - 121786 121674 113 1 2 4 114 130 0.821 9.20
11.02 Intr - 124312 124282 31 0 1 91 72 41 0.140 3.26
11.01 Init - 126465 126398 68 0 2 57 84 58 0.053 3.70
11.00 Prom - 132010 131971 40 0.62
12.00 Prom + 135899 135938 40 3.92
12.01 Init + 138648 138650 3 2 0 80 81 0 0.148 0.76
Suboptimal exons with probability > 1.000
Exnum Type S .Begin ...End .Len Fr Ph B/Ac Do/T CodRg P.... Tscr..
-----
NO EXONS FOUND AT GIVEN PROBABILITY CUTOFF
Predicted peptide sequence(s):
>/tmp/06_03_14-13:05:04.fasta|GENSCAN_predicted_peptide_1|20_aa
MKAIINTEVAVNKRLFGKLL
>/tmp/06_03_14-13:05:04.fasta|GENSCAN_predicted_peptide_2|3_aa
MAN
>/tmp/06_03_14-13:05:04.fasta|GENSCAN_predicted_peptide_3|7_aa
MAKWAVL
>/tmp/06_03_14-13:05:04.fasta|GENSCAN_predicted_peptide_4|5_aa
MTDIS
>/tmp/06_03_14-13:05:04.fasta|GENSCAN_predicted_peptide_5|48_aa
MNPLIAAASVIAAGLAVGLASIGPGVGQGTAAAGQAVEGIARQPEAEK
>/tmp/06_03_14-13:05:04.fasta|GENSCAN_predicted_peptide_6|24_aa
MSTKDFSGDLLQITGMTIEIKNEA
>/tmp/06_03_14-13:05:04.fasta|GENSCAN_predicted_peptide_7|8_aa
MKFKSVFL
>/tmp/06_03_14-13:05:04.fasta|GENSCAN_predicted_peptide_8|14_aa
MKVWQNNYVLELEV
>/tmp/06_03_14-13:05:04.fasta|GENSCAN_predicted_peptide_9|93_aa
MPLMPWATHVLQWAGQRVAISRGNIEITPRTPSCKVKEVGELMTGKPATEAPVNGGRNY
NGPKVAKFLVGRDSDGFSTHTGMRQKYGNRIR
>/tmp/06_03_14-13:05:04.fasta|GENSCAN_predicted_peptide_10|15_aa
MDKREKTIISYQINS
>/tmp/06_03_14-13:05:04.fasta|GENSCAN_predicted_peptide_11|93_aa
MPLMPWATHVLQWAGQRVAISRGNIEITPRTPSCKVKEVGELMTGKPATEAPVNGGRNY
NGPKVAKFLVGRDSDGFSTHTGMRQKYGNRIR
>/tmp/06_03_14-13:05:04.fasta|GENSCAN_predicted_peptide_12|1_aa
M

```

Figura 11.77. Resultados de GENSCAN



BIBLIOGRAFÍA

- ALTSCHUL, S. F.; MADDEN, T. L.; SCHÄFFER, A. A.; ZHANG, J.; ZHANG, Z.; MILLER, W. y LIPMAN, D. J.; *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, Nucleic Acids Res. 25:3389-3402, 1997.
- BAILEY, T. L.; BODÉN, M.; BUSKE, F. A.; FRITH, M.; GRANT, C. E.; CLEMENTI, L.; REN, J.; LI, W. W. y NOBLE, W. S.; *MEME SUITE: tools for motif discovery and searching*, Nucleic Acids Research, 37:W202-W208, 2009.
- BAILEY, T. L. y ELKAN, C.; *Fitting a mixture model by expectation maximization to discover motifs in biopolymers*, Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology Menlo Park, California, AAAI Press, pp. 28-36, 1994.
- BAILEY, T. L. y GRIBSKOV, M.; “Combining evidence using p-values: application to sequence homology searches”, *Bioinformatics*, 14(1):48-54, 1998.
- BAINS, W.; *Ingeniería Genética para todos*, Ed. Alianza Editorial, 1994.
- BAXEVANIS, A. D. y OUELLETTE, B. F. F.; *Bioinformatics. A Practical Guide to the Analysis of Genes and Proteins*, Ed. Wiley & Sons, 2001.
- BESEMER, J.; LOMSADZE, A. y BORODOVSKY, M.; *GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions*, Nucleic Acids Research 29: 2607-2618, 2001.
- BIRNEY, E. y CLAMP, M.; *Biological database design and implementation*, (En) Briefings in Bioinformatics, vol. 5, n.º 1, pp. 31-38, 2004.

- BURGE, C. y KARLIN, S.; *Prediction of complete gene structures in human genomic DNA*, J. Mol. Biol. 268, 78-94, 1997.
- CELMA GIMÉNEZ, M.; CASAMAYOR RÓDENAS, J. C. y MOTA HERRANZ, L.; *Bases de datos relacionales*, Pearson, Prentice Hall, 2003.
- CLAMP, M.; CUFF, J.; SEARLE, S. M. y BARTON, G. J.; "The Jalview Java alignment editor", *Bioinformatics*, 20:426-427, 2004.
- CLAVERIE, J. M. y NOTREDAME, C.; *Bioinformatics for Dummies*, Ed. Wiley Publishing, 2007.
- CONNOLLY, T. y BEGG, C.; *Database Systems. A Practical Approach to Design, Implementation, and Management*, 3.^a Edición, Addison-Wesley, 2002.
- COORAY, M. P. N. S.; "Molecular biological databases: evolutionary history, data modeling, implementation and ethical background", *Sri Lanka Journal of Bio-Medical informatics*, 2012;3(1):2-11.
- DATE, C. J.; *Introducción a los sistemas de bases de datos*, 7.^a Edición, Prentice Hall, 2001.
- FINN, R. D.; BATEMAN, A.; CLEMENTS, J.; COGGILL, P.; EBERHARDT, R. Y.; EDDY, S. R.; HEGER, A.; HETHERINGTON, K.; HOLM, L.; MISTRY, J.; SONNHAMMER, E. L. L.; TATE, J.; PUNTA, M.; *The Pfam protein families database*, Nucleic Acids Research, Database Issue 42:D222-D230, 2014.
- GASTEIGER, E.; HOOGLAND, C.; GATTIKER, A.; DUVAUD, S.; WILKINS, M. R.; APPEL, R. D. y BAIROCH, A.; "Protein Identification and Analysis Tools on the ExPASy Server", (In) John M. Walker (ed): *The Proteomics Protocols Handbook*, Humana Press, pp. 571-607, 2005.
- LEE, M. M.; CHAN, M. y BUNDSCHUH, R.; "Simple is beautiful: a straightforward approach to improve the delineation of true and false positives in PSI-BLAST searches", *Bioinformatics*, n.º 24, pp. 1339-1343, 2008.
- LUSCOMBE, N. M.; GREEBAUM, D. y GERSTEIN, M.; "What is bioinformatics? And introduction and overview", *Yearbook of Medical Informatics*, 2001.
- MATHEWS, C. K., VAN HOLDE, K. E y AHERN, K. G. "Bioquímica". Ed. Addison-Wesley, 2005.
- MORGULIS, A.; COULOURIS, G.; RAYTSELIS, Y.; MADDEN, T. L.; AGARWALA, R. y SCHÄFFER, A. A.; "Database Indexing for Production MegaBLAST Searches", *Bioinformatics*, 24:1757-1764, 2008.

- NAKAI, K. y HORTON, P.; *PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization*, Trends Biochem. Sci, 24(1) 34-35, 1999.
- NELSON, M. R.; REISINGER, S. J. y HENRY, S. G.; "Designing databases to store biological information", *Bioinformatics*, vol. 1, n.º 4, pp. 134-142, 2003.
- PABINGET *et al.*; "A survey of tools for variant analysis of next-generation genome sequencing data", *Briefings in Bioinformatics*, vol. 15, n.º 2, pp. 256-278, 2013.
- PAVLOPOULOS *et al.*; *Unraveling genomic variation from text generation sequencing data*, BioData Mining, 2013.
- PRIMROSE, S. B. y TWYMAN, R. M.; *Principles of Genome Analysis and Genomics*, Ed. Blackwell Science, 2003.
- ROBINSON, T. R.; *Genetics for Dummies*, Ed. Wiley Publishing, 2010.
- ROLDÁN MARTÍNEZ, D.; PASTOR LÓPEZ, O. y VALDERAS ARANDA, P. J.; *Aplicaciones web. Un enfoque aplicado*, Ed. RA-MA, 2009.
- ROLDÁN MARTÍNEZ, D. y VALDERAS ARANDA, P. J.; *Domine Oracle 11g*, Ed. RA-MA, 2013.
- ROLDÁN MARTÍNEZ, D.; VALDERAS ARANDA, P. J. y MARTÍNEZ GÓMEZ, E. J.; *Introducción a Oracle*, Ed. StarBook, 2010.
- SCHATTNER, P.; *Genomes, Browsers and Databases*, Ed. Cambridge University Press, 2008.
- THOMPSON, J. D.; MULLER, A.; WATERHOUSE, A.; PROCTER J.; BARTON, G. J. *et al.*; "MACSIMS: multiple alignment of complete sequences information management system", *BMC Bioinformatics*, 7:318, 2006.
- TROSHIN, P. V.; PROCTER, J. B. y BARTON, G. J.; "Java bioinformatics analysis web services for multiple sequence alignment--JABAWS:MSA", *Bioinformatics*, 27:2001-2002, 2011.
- WATERHOUSE, A. M.; PROCTER, J. B.; MARTIN, D. M. A.; CLAMP, M. y BARTON, G. J.; *Jalview* "Version 2-a multiple sequence alignment editor and analysis workbench", *Bioinformatics*, 25:1189-1191, 2009.
- ZHANG, Z.; SCHWARTZ, S.; WAGNER, L. y MILLER, W.; *A greedy algorithm for aligning DNA sequences*, J Comput Biol, 7(1-2):203-14, 2000.



ÍNDICE ALFABÉTICO

A

Ácido nucleico, 25
Adenina, 24
ADN, 22, 23, 24, 25, 26, 27, 28, 29, 54, 73, 75, 79, 82, 90, 91, 92, 95, 97, 98, 101, 102, 103, 113, 114, 115, 116, 117, 119, 132, 135, 137, 139, 140, 160, 185, 195, 196, 197, 215, 231, 237, 249
Alineamiento de pares, 117, 118, 119, 160, 185
Alineamiento global, 116, 185
Alineamiento local, 116
Alineamiento múltiple, 94, 117, 119, 120, 166, 174, 185, 190, 193
Aminoácidos, 26
Análisis de secuencias, 82, 113, 122
Anotación, 11, 31, 35, 63, 83, 126, 127, 205
ARN, 19, 20, 21, 24, 25, 26, 27, 28, 56, 73, 74, 82, 92, 116, 132, 144, 169, 177, 196

B

BAM, 31, 36, 38, 40
Base de datos, 53, 54, 56, 57, 58, 59, 60, 61, 62, 65, 67, 74, 75, 78, 79,

82, 84, 87, 89, 91, 93, 94, 95, 96, 97, 102, 109, 118, 119, 120, 129, 139, 150, 156, 160, 161, 163, 164, 165, 166, 181, 201, 202, 203, 204, 205, 206, 211, 217, 218, 219, 220, 221, 235, 237, 248
Bases de datos genómicas, 16, 51, 54, 55, 57, 58, 73, 90, 125
BED, 31, 48, 49
Bioinformática, 15, 16, 17, 32, 97, 167
Biología molecular, 16, 27, 113, 215
Biosíntesis, 21
Biotecnología, 16
BLAST, 94, 118, 119, 160, 161, 162, 163, 164, 179, 180, 181, 182, 198, 201, 202, 203, 217, 255, 256

C

Citosina, 32, 185, 186, 187, 188
Clustal, 120
Código genético, 25, 26, 102
Codón, 26, 42, 102, 114, 124, 150, 154, 239
Consenso, 94, 117, 178
Cromosoma, 22, 23, 43, 44, 45, 48, 54, 110, 122, 123, 124

D

Datos biológicos, 15, 53, 57, 126

E

EMBL, 41, 42, 55, 75, 91, 94
Entrez, 53, 55, 98, 106, 108, 109
Estructura primaria, 197, 209, 237, 239
Estructura secundaria, 89, 197, 221, 242, 244, 245
Estructura terciaria, 197, 212, 247
Eucariotas, 20, 22, 23, 27, 29, 51, 82, 103, 113

F

FASTA, 31, 34, 35, 82, 99, 100, 130, 157, 161, 167, 180, 186, 187, 192, 217, 228, 249
FASTAQ, 31, 35
Fenotipo, 22, 95, 96, 122
Filogenia, 185, 189

G

GenBank, 51, 52, 55, 60, 73, 74, 75, 78, 80, 81, 82, 83, 91, 92, 93, 97, 99, 100, 101, 102, 103, 106, 107, 109, 150, 153, 154, 156, 157, 182
Genética, 19, 90, 255
Genotipo, 44
GFF, 31, 41
GFF3, 31, 41, 42, 43
Gráfico de puntos, 117, 118, 139, 141, 142, 143, 144
Guanina, 32
GVF, 31, 43, 44, 45

I

Información biológica, 15, 53, 56, 59, 75, 79, 81, 126
Ingeniería biomédica, 16
Integración, 11, 60, 61, 62

J

Jalview, 11, 167, 168, 174, 176, 177, 178, 191, 255, 257

M

Matriz de sustitución, 121, 193
Modelado, 56, 57, 59, 113, 115, 119, 123, 197, 198, 216, 247
Motivos funcionales, 185, 201

N

NCBI, 44, 51, 53, 74, 75, 80, 82, 83, 91, 92, 95, 98, 99, 103, 110, 129, 150, 159, 161, 202, 249

O

ORF, 114, 119, 150, 151, 152, 153, 154, 159

P

PDB, 55, 76, 82, 84, 85, 86, 87, 88, 89, 104, 217, 219
Procariotas, 20, 22, 27, 29, 51, 97, 102, 114, 214
Promotor, 102
Proteoma, 214

R

Replicación, 28

S

SAM, 31, 36, 37, 38, 39
Secuenciación, 31, 35, 36, 37, 51, 55, 96, 113, 114, 115, 116, 126, 129
Secuencia de bases, 24, 82, 152
Secuencia de nucleótidos, 28, 32, 75, 80, 97, 101, 109, 119, 140, 155, 158, 160, 186
Secuencia proteica, 113, 119, 166, 208, 216, 228, 237, 240, 243

T

Traducción, 33, 75, 107, 124, 151,
237, 238

U

UniProt, 55, 84, 215
Uracilo, 24

V

Variación, 43, 44, 45, 46, 48, 65, 96,
106, 107, 122, 124, 125
VCF, 31, 45, 46, 47

X

XML, 59, 67, 68, 70

